TU WIEN Informatics

Social Media User Profiling for Credit Scoring: A Taxonomy of Explainability Techniques

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering/Internet Computing

eingereicht von

Valdemar Lipenko
00627631

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuer: Thomas Grechenig

Wien, 05.07.2021

Unterschrift Verfasser | Unterschrift Betreuer

Technische Universität Wien
Karlsplatz 13 | 1040 Wien | +43-1-58801-0 | www.tuwien.at

TU WIEN Informatics

Social Media User Profiling for Credit Scoring: A Taxonomy of Explainability Techniques

MASTER'S THESIS

submitted in partial fulfilment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering/Internet Computing

by

Valdemar Lipenko
00627631

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Thomas Grechenig

Vienna, 05.07.2021

Signature Author | Signature Advisor

Technische Universität Wien
Karlsplatz 13 | 1040 Wien | +43-1-58801-0 | www.tuwien.at

Social Media User Profiling for Credit Scoring: A Taxonomy of Explainability Techniques

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering/Internet Computing

eingereicht von

Valdemar Lipenko

00627631

ausgeführt am

Institut für Information Systems Engineering

Forschungsbereich Business Informatics

Forschungsgruppe Industrielle Software

der Fakultät für Informatik der Technischen Universität Wien

Betreuung: Thomas Grechenig

Technische Universität Wien, Forschungsgruppe INSO

A-1040 Wien ▪ Wiedner Hauptstr. 76/2/2 ▪ Tel. +43-1-587 21 97 ▪ www.inso.tuwien.ac.at

Erklärung zur Verfassung der Arbeit

Valdemar Lipenko
Löhrgasse 18/38, 1150 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

----------	----------
(Ort, Datum)	(Unterschrift Verfasser)

Kurzfassung

Die Schaffung inklusiver Finanzdienstleistungen, um insbesondere derzeit ausgeschlossenen Personengruppen den Zugang zu Verbraucherkrediten ermöglichen zu können, macht zusätzliche Informationsquellen zur Durchführung von Kreditwürdigkeitsprüfungen notwendig. Die umfassenden Daten über die Nutzer der mittlerweile weltweit verbreiteten Social Media-Plattformen sind somit jene Daten, die zu diesem Zweck verwendbar sein könnten. Eine der besonderen Herausforderungen besteht darin, die Erklärbarkeit der möglichen Ansätze von Social Media Profiling, die zwecks Credit Scoring eingesetzt werden könnten, sicherzustellen. Neueste Forschungsergebnisse haben dabei die unterschiedlichsten Erklärbarkeitstechniken für die Machine Learning Ansätze bereits aufgezeigt. Es fehlt jedoch an umfassender Zuordnung dieser Erklärbarkeitstechniken zu genau den Ansätzen, die potentiell die Komponenten der Credit Scoring Modelle aus Social Media Daten ableiten könnten. Ziel dieser Arbeit ist es daher, eine Taxonomie von Erklärbarkeitstechniken für Social Media Profiling Ansätze zu erstellen, die zwecks Credit Scoring eingesetzt werden könnten. Zur Erreichung dieses Ziels wurde die Methodologie zur Entwicklung von Taxonomien in Software Engineering befolgt. Die erste Phase umfasst die Planung der Taxonomie mit der Definition vom Kontext und der Angabe der Hauptaspekte der zu entwickelnden Taxonomie. In der zweiten Phase findet die Identifizierung von Begriffen der Taxonomie mithilfe des systematischen Literaturreviews statt. Die Erstellung der Taxonomie erfolgt in der dritten Phase durch die Kategorisierung der identifizierten Begriffe und die Feststellung der Beziehungen zwischen den Kategorien. Die Expertenbefragung ist in der abschließenden vierten Phase zur Validierung der Taxonomie eingesetzt. Das Ziel der Arbeit wurde erfolgreich erreicht. Die erstellte Taxonomie deckt 496 Komponenten der Credit Scoring Modelle, 574 Social Media Profiling Ansätze und 640 Erklärbarkeitstechniken ab. Auf der Ebene der Komponenten der Credit Scoring Modelle sind sowohl die gut erforschten (z.B. die Bonitätsgeschichte, die demografische Daten und das Beschäftigungsverhältnis) als auch sehr spezielle Kategorien (z.B. Look-a-likes und der potentielle Einfluss psychologischer Variablen) erfasst. Fast alle dieser Kategorien sind durch identifizierte Social Media Profiling Ansätze ableitbar. Die Ausnahmen sind nachvollziehbar. So benötigen die Attribute vom beantragten Kredit oder die Daten über die Geschäftsbeziehung zwischen Kreditgeber und Verbraucher beispielsweise keine zusätzliche Ableitung. Für die meisten Kategorien der Social Media Profiling Ansätze sind Erklärbarkeitstechniken verfügbar, bis auf Dimensionality Reduction, Social Semantic Web und Algorithmen aus der Graphentheorie, für die keine anwendbaren Erklärbarkeitstechniken identifiziert wurden. Die erstellte Taxonomie trägt zu einem besseren Verständnis der verfügbaren Erklärbarkeitstechniken für Ansätze bei, mit denen potenziell Komponenten der Credit Scoring Modelle aus Social Media Daten abgeleitet werden können. Die erstellte Taxonomie wurde erfolgreich validiert, indem die Expertenmeinung klassifiziert wurde.

Keywords: *explainability techniques, social media user profiling, credit scoring, taxonomy*.

Abstract

The aim to enable more inclusive financial services, particularly to improve the access to consumer credits, leads to the discovery of additional sources of information to conduct credit scoring. At the same time, the recent expansion of social media, which contains valuable information from billions of people around the world, is tremendous. Thus, social media data is naturally a potential candidate to be part of a solution for improved consumer credit offering. Among different requirements around possible applications of social media user profiling approaches to derive credit scoring model components, one that is particularly challenging is to ensure the explainability of such approaches. On the one side, recent research contributed various explainability techniques to modern machine learning approaches. On the other side, there is a lack of concrete mapping between these explainability techniques and the social media user profiling approaches potentially capable of deriving credit scoring model components. Hence, the aim of this work is to construct a taxonomy of explainability techniques for social media user profiling approaches in credit scoring. To achieve this goal, the methodology for developing taxonomies in software engineering is followed. The first phase is the planning phase, with the specification of the context and the defining aspects of the taxonomy. Extraction of the relevant terms is performed in the second phase by systematic literature reviews. The third phase covers taxonomy design and construction through categorization of the identified terms and establishment of the relationships between them. Experts' opinion survey is conducted for the validation of the developed taxonomy in the final fourth phase. The aim of the thesis has been successfully achieved. The constructed taxonomy covers 496 credit scoring model components, 574 social media user profiling approaches, and 640 explainability techniques. On the level of credit scoring model components well researched (such as credit history, demographic data, and employment status) and more specific categories (such as look-a-likes and potential influence of various psychological variables) are captured. Almost all of the categories of credit scoring model components are potentially derivable by the identified social media profiling approaches. The few exceptions are justified (e.g., data on bank-borrower relationship or attributes of credit applied for do not require to be explicitly derived). For almost all of the categories of social media user profiling approaches there are various explainability techniques available, with the exception of dimensionality reduction, graph theory algorithms, and social semantic web, for which no evidence of available explainability techniques found. The developed taxonomy contributes to improved understanding of currently available explainability techniques of user profiling approaches applicable to potentially derive credit scoring model components from social media data. The constructed taxonomy is successfully validated by classifying experts' opinions.

Keywords: *explainability techniques, social media user profiling, credit scoring, taxonomy*.

Table of contents

List of figures

List of tables

1 Introduction

1.1 Motivation

Social media is clearly omnipresent in modern society, with worldwide spread and the amount of constantly produced data reaching previously unthinkable dimensions. Facebook, which is the largest social network in the world (Global social media ranking 2019 | Statista, 2020), counts currently more than 2 billion active users (Global social media ranking 2019 | Statista, 2020). Besides Facebook, there are about two dozen of other social network sites with over 100 million users (Global social media ranking 2019 | Statista, 2020). One of the main contributors to this success is undoubtedly the social media data itself, which usually consists of users' service-specific profiles, user-generated content (such as submitted photos, posts, tags, comments, likes, etc.), and a large number of connections (e.g., between users, to specific groups or topics) resulting in impressively complex networks (Kaplan & Haenlein, 2010; Boyd & Ellison, 2010). The speed with which social network services are expanding is very remarkable, leading to a large controversy and many discussions regarding potential privacy threats posed by various applications of utilizing social media data. Since in many cases it is not reasonable to just abandon the use of social media data, the following aspects are coming to the fore: for which use cases and how can social media data be successfully utilized in a privacy-preserving manner, whether legal systems are capable to keep pace with recent developments, what is acceptance in the population of using their social media data for those particular use cases.

The most straightforward component to generate profit for social network sites is to offer online advertisement. The deeper information that social network services are able to collect about their users the better targeted social advertising on their sites can be offered (Bakshy, Dean, Rong, & Itamar, 2012). Third parties are often also interested in accessing social media data with other than marketing intentions. One example is to conduct pre-employment screening based on the data from social network services (Ebnet, 2012; Stoughton, Thompson, & Meade, 2015). Furthermore, it is possible to harness social media for disaster relief, in cases of emergency and catastrophes (Crawford & Finn, 2015; Gao, Barbier, & Goolsby , 2011), or even using social network sites for predicting depression and suicide numbers (De Choudhury, Gamon, Counts, & Horvitz, 2013; Won, et al., 2013). There are also companies that are building their business models completely based on social media data, in particular those offering consumer loans: the

key component of Kreditech from Germany, Lenddo from Singapore, or Social Lender from Nigeria is to assess the consumers' creditworthiness taking into account also their social media data (Cullerton, 2012; Alpar, 2016; Packin & Lev-Aretz, 2016).

Probably the most comprehensive attempt to make use of social media data is currently undertaken in China with the Social Credit System (SCS), which is intended to become mandatory for every Chinese citizen and business entities (Chen & Cheung, 2017; Kshetri, 2016). Stated aim is to encourage trustworthiness in complying with legal rules, moral norms, and professional standards, at the same time punishing untrustworthiness (Chen & Cheung, 2017). Economic behaviour and compliance with ethical standards should become interconnected (Kshetri, 2016). To achieve this goal, the Chinese government intends to observe and to evaluate social behaviour of its citizens in addition to their financial activities, criminal record, etc. (Kshetri, 2016). Hence, also integration of social media data (such as posts, likes, comments, but also connections, up to online search, and other personal data) is planned to contribute to the overall ranking process (Chen & Cheung, 2017; Kshetri, 2016; Han, 2017; Diab, 2017).

Following the active developments to consider social media data for credit scoring, an important additional requirement to such undertakings is to ensure explainability of applicable technical approaches, with various explainability techniques being the main focus of this work.

1.2 Problem statement

Financial service providers active in the field of credit scoring strive for predictive models that are both accurate and explainable in order to use these models in practice. The explainability is even more important when innovative data sources such as social media are used for financial purposes. The currently very dynamic field of models explainability result in a wide range of different techniques to achieve the desired explainability. This leads to the problem of selecting adequate explainability techniques for the decisions regarding social media profiling use in credit scoring, namely lack of a classification (e.g., in the form of a taxonomy) of such explainability techniques. Hence, the focus of this diploma thesis is on the following two research questions.

RQ1. What are the techniques to provide explainability of social media user profiling approaches in credit scoring?

RQ2. What are the valid relationships in the taxonomy of explainability techniques for social media profiling in credit scoring?

1.3 Aim of the work

The expected result is a classification of techniques for explainability of social media profiling in credit scoring in taxonomy form. The validity of the developed taxonomy is to be ensured through experts' opinion survey.

The main target audience interested in the expected outcomes are financial service providers that offer consumer credits, such as banks, credit card providers, and other institutions involved in consumers lending business. A validated taxonomy of explainability techniques for social media profiling in credit scoring is highly important for assessing the potential and challenges of implementing social media profiling for credit scoring, in particular its conformity with legal requirements of explainability. A minor additional target group is researchers in the field of social media profiling, explainable machine learning.

Furthermore, for a such rather controversial undertaking as using social media data in credit scoring the evaluation of possible ethical issues is an important contribution to sensibilize and raise awareness.

1.4 Methodology

Taxonomy of explainability techniques for social media user profiling approaches in credit scoring is developed following the methodology for developing taxonomies in software engineering (Usman, Britto, Börstler, & Mendes, 2017). The single phases with the corresponding activities conducted in each phase are provided in Table 1.

Phase	Activities
Planning	[A01] Define SE knowledge area [A02] Describe the objectives of the taxonomy [A03] Describe the subject matter to be classified [A04] Select classification structure type [A05] Select classification procedure type [A06] Identify the sources of information

Identification and extraction	[A07] Extract all terms [A08] Perform terminology control
Design and construction	[A09] Identify and describe taxonomy dimensions [A10] Identify and describe categories of each dimension [A11] Identify and describe the relationships [A12] Define the guidelines for using and updating the taxonomy
Testing and validation	[A13] Validate the taxonomy

Table 1. Taxonomy development methodology (Usman, Britto, Börstler, & Mendes, 2017)

The single stages of the methodological approach are as follows.

- Systematic literature review (SLR) (Kitchenham & Charters, 2007) for state-of-the-art analysis, i.e., comparable taxonomies of explainability techniques in the field of user profiling, social media profiling, or machine learning in general. State of the art is to be described in chapter 2.
- Evaluation of systems that consider social media for credit scoring, outline of legal side from the point of view of the GDPR, and elaboration of the potential ethical issues following the approach to assess the ethical implications of data science elaborated by (Floridi & Taddeo, 2016), which are respectively to be discussed in chapter 3.
- Taxonomy planning phase (Usman, Britto, Börstler, & Mendes, 2017) contains the defining aspects of the taxonomy, such as software engineering knowledge area that it is associated with, objectives and subject matter of the taxonomy, taxonomy structure and procedure types, identifies information sources for taxonomy development. Taxonomy planning is to be provided in chapter 4.
- Extraction of relevant terms and terminology control (Usman, Britto, Börstler, & Mendes, 2017) are conducted by systematic literature reviews (SLR) (Kitchenham & Charters, 2007) of components affecting credit scores of consumer credit applicants, approaches for social media user profiling applicable in credit scoring, and explainability techniques for social media user profiling approaches applicable in credit scoring. Taxonomy terms identification is to be provided in chapter 5.
- Taxonomy design and construction (Usman, Britto, Börstler, & Mendes, 2017) contain elaborations on taxonomy dimensions, categories of dimensions with relationships between them, and guidelines for using and updating the taxonomy. Taxonomy construction is to be provided in chapter 6.

- Taxonomy validation by qualitative cross-sectional experts' opinion survey using non-probabilistic convenience sampling (Cresswell, 2012) through expert interviews in adoption of user profiling specific approaches in financial services domain. The exact criteria of survey potential participants are to be appropriately considered, focusing on those experienced in statistical modelling and/or machine learning. The access to the respective experts is to occur through the connections in the Balancing Banks division of the research group Industrial Software at the Vienna University of Technology, personal contacts, and by the professional networking capabilities on the social network LinkedIn. Taxonomy validation is to be provided in chapter 7.

2 State of the Art

State of the art review proceeds by applying the systematic literature review (SLR) (Kitchenham & Charters, 2007). The focus is on the following research questions:

CH2-SLR-RQ1. What are the research activities to study explainability of social media user profiling for credit scoring in the recent years (from the 1st of January 2015 until the 1st of January 2020)?

CH2-SLR-RQ2. Which most significant taxonomies related to explainability of social media user profiling for credit scoring exist?

CH2-SLR-RQ3. What is the experts' opinion on the most significant taxonomies related to explainability of social media user profiling for credit scoring exist?

CH2-SLR-RQ4. What are the limitations of the current research on explainability of social media user profiling for credit scoring and the main challenges that are to be addressed in the future research?

The search process, the inclusion and the exclusion criteria, the data collection, and the data analysis of the conducted SLR are explained as next. After that, the discussions to the defined in this chapter research questions are provided.

The search process is a manual search of suitable academic publications and literature using CatalogPlus, the comprehensive TU Wien academic research portal (TU CatalogPlus Search, 2020). The search string construction is based on the following considerations regarding the main terms of interest, taking into account the search capabilities of CatalogPlus (e.g., the limitation to three search fields).

- First of all, the focus is on "social media" and "profiling". The closest synonym to "social media" is "social networks", which also receives wide coverage in academic publications. At the same time, it would be too limiting to include, for example, "social network sites", "online social networks" or other more specific terms.
- To receive as many search results as feasible relevant to "credit scoring", the best decision is to consider solely "credit", hence covering also such terms as "credit score", "credit analysis", "credit evaluation", etc. (even with the resulting minor disadvantage of preliminary having to deal with more irrelevant search results to go through). An explicit inclusion of "loan" is believed to be unnecessary, since publications using term "loan"

would most probably also contain "credit" in at least some of the fields, and the focus is moreover on "credit".

- "Explainability" is more limiting than e.g., "explainable" models. At the same time, "interpretability", which is sometimes considered as a synonym to "explainability", is the term semantically used to describe slightly different characteristics, hence not of an interest in the current context.
- Eventually, to construct a "taxonomy" is the core interest of this thesis, hence necessarily covered. A legitim and most straightforward synonym with almost completely the same meaning as "taxonomy" is "classification", but resulting in a major disadvantage of having to deal with unfeasibly large amount of additional and irrelevant search results to preliminary having to go through.
- In order to account for unfeasibility to add "classification" as a search term to each of the search queries, but still covering this important to "taxonomy" synonym, as an addition "classification" is beneficiary to be used in conjunction with "explainable", same as also "taxonomy" together with "explainable". Although the number of the search results would be relatively large, it is ensured to completely cover topics of interest.
- For providing an answer to the research question CH2-SLR-RQ3 regarding validation of the related taxonomies (through experts' opinion survey) the decision is made not to contain that as a separate search term, but in all of the selected publications in the evaluation stage to analyse how the validation was achieved, whether explicit validation (experts' opinion survey) was conducted.

The final search string is consequently as follows:

(((“social media” OR “social networks”) AND “profiling”) AND (“credit” OR “explainable” OR “taxonomy”)) OR (“explainable” AND (“taxonomy” OR “classification”))

Table 2. State of the art SLR search string

The search string is accordingly customized to be used in CatalogPlus: split to at most three search fields for each search query; use exact match for single terms; conduct search through all fields of publications.

The inclusion of only results that are expected to help to address the specified research questions is ensured through the following inclusion criteria.

CH2-SLR-IC1. Published between the 1st of January 2015 and the 1st of January 2020.

CH2-SLR-IC2. Published in a peer-reviewed journal.

CH2-SLR-IC3. Focus on social media profiling.
CH2-SLR-IC4. Explainability or applicability in credit scoring explicitly addressed.
CH2-SLR-IC5. Taxonomy of techniques for explainability of approaches related to social media profiling addressed.

The following single exclusion criterion is taken into account during publications selection stage.

CH2-SLR-EC1. There is no access through TU Wien student account.

During the data collection stage, a search protocol was used, the summary of which is depicted in the Table 3. The inclusion criteria CH2-SLR-IC1, CH2-SLR-IC2, CH2-SLR-IC3, CH2-SLR-IC4 were applied to the part of the search string ((("social media" OR "social networks") AND "profiling") AND ("credit" OR "explainable" OR "taxonomy")), and the inclusion criteria CH2-SLR-IC1, CH2-SLR-IC2, CH2-SLR-IC5 to the part of the search string ("explainable" AND ("taxonomy" OR "classification")).

Search query	Results	IC1	IC1, IC2	IC1, IC2, IC3	IC1, IC2, IC3, IC4	IC1, IC2, IC5	EC1	Selected
"social media" AND "profiling" AND "credit"	2355	1587	370	31	4	-	0	4
"social networks" AND "profiling" AND "credit"	958	535	249	21	4	-	0	4
"social media" AND "profiling" AND "explainable"	48	38	18	2	0	-	0	0
"social networks" AND "profiling" AND "explainable"	43	25	13	3	1	-	0	1

"social media" AND "profiling" AND "taxonomy"	455	287	178	18	0	-	0	0
"social networks" AND "profiling" AND "taxonomy"	495	246	149	13	1	-	0	1
"explainable" AND "taxonomy"	755	205	132	-	-	2	0	2
"explainable" AND "classification"	3183	1035	735	-	-	5	0	5

Table 3. State of the art SLR data collection

From each selected publication its bibliographic information (title, authors, publication year) and main topic areas (e.g., social media profiling, explainability) are extracted. The obtained data (without duplicates) is tabulated to show the extracted information, as Table 4 illustrates. The content of each publication is then analysed to provide the answers to the defined chapter research questions.

Title	Author(s)	Year	Main Topic(s)
Not-So-Big and Big Credit Data Between Traditional Consumer Finance, FinTechs, and the Banking Union: Old and New Challenges in an Enduring EU Policy and Legal Conundrum	Ferretti, F.	2018	consumer data for creditworthiness assessment in the EU
Personal credit profiling via latent user behavior dimensions on social media	Guo, G.; Zhu, F.; Chen, E.; Wu, L.; Liu, Q.; Liu, Y.; Qiu, M.	2016	credit profiling using user behavior analysis from online social data
From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring	Guo, G.; Zhu, F.; Chen, E.; Liu, Q.; Wu, L.; Guan, C.	2016	social media data mining for credit scoring

Big Data-Scoring unter dem Einfluss der Datenschutz-Grundverordnung	Eschholz, S.	2017	social media data for credit scoring under the GDPR
European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"	Goodman, B.; Flaxman, S.	2017	algorithmic decision making under the GDPR
A Survey of Methods for Explaining Black Box Models	Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D.	2019	classification of problems of explaining decision support systems
A systematic review and taxonomy of explanations in decision support and recommender systems	Nunes, I.; Jannach, D.	2017	taxonomy of explanations in decision support
Toward Human-Understandable, Explainable AI	Hagras, H.	2018	overview and introduction to explainable AI systems
Defining Explainable AI for Requirements Analysis	Sheh, R.; Monteath, I.	2018	categorization of requirements to explainable AI
Increasing Transparency in Algorithmic-Decision-Making with Explainable AI	Waltl, B.; Vogl, R.	2018	levels of transparency in algorithmic decision making

Table 4. State of the art SLR search results

Following is the discussion about the findings to answer stated chapter research questions.

CH2-SLR-RQ1. What are the research activities to study explainability of social media user profiling for credit scoring in the recent years (from the 1st of January 2015 until the 1st of January 2020)?

The amount of the selected publications and their distribution by year of publication results in no particular trend observable in the researched recent five years. At the same time, there is a clear interest in the research of such specific topic as using social media for credit scoring (Ferretti, 2018; Guo, et al., 2016; Guo, et al., 2016; Eschholz, 2017). A number of publications focus in particular on the legal side of social media profiling, specifically under the consideration of the EU's GDPR, most prominently due to the novel right to explanation (Ferretti, 2018; Eschholz, 2017; Goodman & Flaxman, 2017). The overall issue of explainability, which especially recently gains on popularity in the academic publications, is still either covered in general for the complete field of explainable AI (Hagras, 2018; Sheh & Monteath, 2018), or partly already for some specific areas such as decision support systems (Guidotti, et al., 2019; Nunes & Jannach, 2017; Waltl & Vogl, 2018), when relating to social media profiling. In particular, Nunes & Jannach (2017) provide taxonomy of explanations and Guidotti, et al. (2019) provide classification of problems of explanations in decision support systems, Sheh & Monteath (2018) focus on categorization of requirements to explainable AI.

> CH2-SLR-RQ2. Which most significant taxonomies related to explainability of social media user profiling for credit scoring exist?

A number of recent publications focus on elaborating explainability-related classifications in the field of decision support systems, where used approaches are somewhat similar to those used in social media profiling. Nunes & Jannach (2017) conducted a systematic review to develop a taxonomy of explanations in decision support systems. As the result of investigating the purposes of explanations, the different techniques to generate, to present to users and to evaluate explanations a comprehensive taxonomy of explanation aspects is derived, which are to be considered when designing the explanation facilities for advice-giving systems, as depicted on the Figure 1.

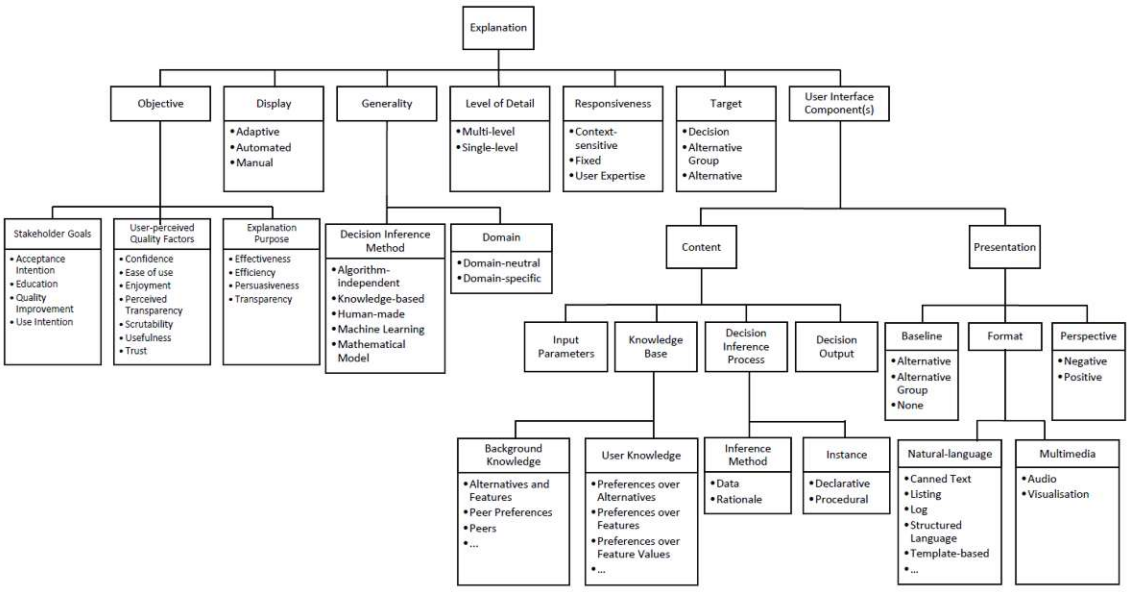

Figure 1. Taxonomy of explanation aspects in decision support systems (Nunes & Jannach, 2017, p. 33)

Waltl & Vogl (2018) also focus on the decision support systems, namely the issue of transparency assessment in algorithmic decision-making. Their brief investigation proceeds by dividing the assessment of transparency of algorithmic decision-making systems into the following three levels: transparency of the process of decision-making system development, transparency of the model with decision-making structure, and transparency of the concrete classification instance.

Guidotti, et al. (2019) conducted another explainability-related survey in the field of decision support systems. Their focus is on the problems with respect to the notion of explanation in decision support systems, providing a comprehensive classification thereof. On the top level, the distinction between the following four categories of problems is proposed: model explanation problem, outcome explanation problem, model inspection problem, and transparent model design problem. On the next level, categories of solutions proposed in the surveyed publications to develop concrete explanators to decision support systems are described, which are: decision tree, decision rules, features importance, salient mask, sensitivity analysis, partial dependence plot, prototype selection, neurons activation.

Sheh & Monteath (2018) focus on categorization of requirements generally to AI systems in order for them to be perceived as trustworthy. The following three dimensions for the categorization of explanatory requirements are proposed: source of the explanation (either from the AI system itself or from another system observing the overall process), depth of the explanation (per attribute versus complete model), explanation scope (either justification of a concrete decision or teaching to understand e.g., also similar decisions). Subsequently, the capabilities of the selected ML

techniques to provide explanatory requirements based on the proposed categorization are investigated in details.

> CH2-SLR-RQ3. What is the experts' opinion on the most significant taxonomies related to explainability of social media user profiling for credit scoring exist?

The most significant taxonomies related to explainability of social media profiling for credit scoring, which were described in the discussion to CH2-SLR-RQ2, do not contain explicit experts' opinion surveys to ensure their validity. On the one hand, the high significance of these academic publications is justified by the process of academic publications being published in the peer-reviewed journals. On the other hand, either the validity is stated to be ensured by chosen methodology (structured literature review with a subsequent accurate analysis as by Nunes & Jannach (2017) and Guidotti, et al. (2019)) or the issue of validity is not addressed (Sheh & Monteath, 2018; Waltl & Vogl, 2018).

> CH2-SLR-RQ4. What are the limitations of the current research on explainability of social media user profiling for credit scoring and the main challenges that are to be addressed in the future research?

The academic publications assessed in the discussion to CH2-SLR-RQ2 possess the following limitations regarding explainability techniques for social media profiling in credit scoring. Nunes & Jannach (2017) developed a comprehensive taxonomy of the explanations overall, without mentioning concrete explanation techniques, and also covering neither social media profiling nor credit scoring in particular. Waltl & Vogl (2018) provide a brief overview only of the transparency issue of XAI, and also without providing a comprehensive classification. Guidotti, et al. (2019) focus on the selected decision support systems' black box model types, providing the classification by problems, hence not a comprehensive view on the concrete explainability techniques specifically for social media profiling. Sheh & Monteath (2018) categorize the overall requirements to XAI, with the resulting limitations for the current context.

As the result of the conducted structured review and the subsequent comprehensive analysis of selected publications, Guidotti, et al. (2019) also identified some of the currently open research questions and future research directions. Main issue is lack of a common agreement on the exact meaning of "explanation" for black box models, with different works providing as "explanation" e.g., set of rules, decision trees, prototypes, etc. Regarding the concrete desired properties for an explanation to possess, in particular "*no work that seriously addresses the problem of quantifying*

the grade of comprehensibility of an explanation for humans" (Guidotti, et al., 2019, p. 37) is known to exist.

To summarize, a comprehensive overview of the techniques used in social media profiling for credit scoring is missing, same as explainability considerations specifically of social media profiling for credit scoring. As the result, there is also lack of classifications of techniques to achieve explainability of social media profiling for credit scoring. Hence, the main aim of this thesis to provide a validated taxonomy of the explainability techniques for social media profiling in credit scoring is a very important contribution to the current research in this field.

3 Social media data for credit scoring

The controversial undertaking to utilize data from social media for the purposes of conducting credit scoring touches different aspects, ranging from legal background to potential ethical issues. Hence, the elaboration of the following subsections aims to facilitate better understanding of this thesis' complex background of using social media data for credit scoring.

The subsection 3.1 provides an overview of the most prominent systems that are known to consider social media data for credit scoring. The subsection 3.2 covers legal aspects from the perspective of Austria, i.e., the most relevant GDPR clauses applicable in the present context. The subsection 3.3 outlines the potential ethical issues resulting from using social media data for credit scoring.

3.1 Systems using social media for credit scoring

There are quite some examples around the world that put into practice the idea to consider social media data for credit scoring. The evaluation of the top 50 search results on Google (that accounts for ca. 92% of search engines market share worldwide (Search Engine Market Share Worldwide | StatCounter Global Stats, 2020)) for the search string *"social media" AND "credit scoring"* led to the following systems identified (in alphabetic order): Accion, Affirm, Alipay, Big Data Scoring, Brigit, Crediograph, Creditinfo, CredoLab, Demyst Data, Earnest, FriendlyScore, Hello Soda, Kabbage, Kiva, Kreditech, Lenddo, Line Score, Lodex, Moven, NeoVerify, Oportun, Petal, SOCSCOR, Social Credit System, Tala, WePay, Wonga, ZestFinance. The total number of the mentions of each of these systems is depicted on the Figure 2, with Lenddo, Kreditech, or Social Credit System mentioned in ca. 90% of the cases when at least one concrete system is mentioned. Hence, these 3 systems are considered as currently the most prominent cases of considering social media data for credit scoring, and they are outlined in the following subsections in details.

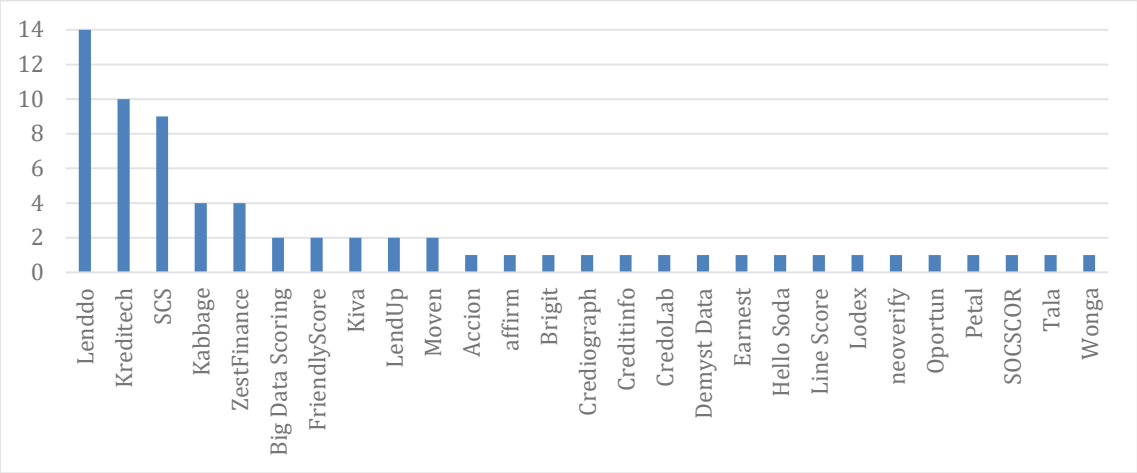

Figure 2. Systems potentially using social media data in credit scoring by the total number of their mentions in the top 50 search results on Google search to *"social media" AND "credit scoring"*

3.1.1 Lenddo

Lenddo (or LenddoEFL) reached probably the largest expansion among technology companies operating in the credit analysis domain based on non-traditional data, in particular data available from social media. Lenddo covers by far more countries and larger population than another major system (About Lenddo, 2018). Nevertheless, Lenddo underwent a remarkable development from providing loans based on internal algorithms to completely abandoning this business model after four successful years of operation in favour of providing the access to their algorithms to external companies from different sectors instead (About Lenddo, 2018) (Packin & Lev-Aretz, 2016, p. 365) (Costa, Deb, & Kubzansky, 2015, p. 56). Hence, (Packin & Lev-Aretz, 2016) derive from this a conclusion that the value of credit scoring algorithms and gained knowledge apparently exceeds the value of the lending business itself (p. 365).

Lenddo focuses mainly on social media activities of consumers for its credit score computation algorithm (Costa, Deb, & Kubzansky, 2015, p. 56). The target group is the emerging middle class in the developing countries, whereas concentrating again primarily on short-term microloans (About Lenddo, 2018). First, users are granting permission to access their social media profiles, such as on Facebook, LinkedIn, Google, Yahoo and Twitter, in addition to optionally providing also other non-traditional data. Lenddo extends this data then with the available traditional credit scoring data on that particular customer in order to make the most accurate and precise decision regarding the possibility for the applicant to get a credit, and if so, under which particular conditions (Packin & Lev-Aretz, 2016, p. 361). The handling of consumers' connections by Lenddo is especially noteworthy: not only that the character of the candidate is precisely studied

under the consideration of his social network, additionally the credit scores of his connections are badly affected in case of default (Packin & Lev-Aretz, 2016, p. 361).

After assessing hundreds of thousands of loan applications in the first years of operations, Lenddo now offers the product called LenddoScore (Our Products, 2018) to companies within and outside of the financial sector. The patented score ranges from 1 to 1000 and serves as an estimate of the likelihood of default based on the prediction of an individual's character (the higher score represents the higher probability of default). For this purpose, various machine learning techniques are in place, which, among others, potentially produce new predictive features. It is stated that less than 3 minutes are required in order to provide a score to a particular request. The LenddoScore completely relies on non-traditional data, such as telecom data, mobile data, browser data, data from social networks and from e-commerce, financial transactions data, form filling analytics, and psychometric data. Nevertheless, Lenddo does not aim to completely replace the traditional underwriting tools with the offered score, but to complement them. At the same time, the stated goal is to facilitate a higher number of approved applications together with improving profitability and reducing risk.

(Costa, Deb, & Kubzansky, 2015) found out that consumers in the countries, where Lenddo operates, are largely (e.g., 70% of the Colombian consumers) willing to share their non-traditional data, in particular social media data and web browsing history, in order to get the chance of improving their creditworthiness and potential conditions for a loan. Similarly, in Tanzania the need for a credit also usually supersedes privacy concerns (Costa, Deb, & Kubzansky, 2015, p. 56). Hence, it should be aimed by respective data providers to accurately implement the required precautions in order to protect the privacy of own customers in particular in cases, where they might neglect its importance on their own. In case of Lenddo, there is also additional implication of where its score might be used, stating to offer the possibility to not only unlock loans, but also potential improving chances of employment (About Lenddo, 2018), or, perhaps, vice versa.

3.1.2 Kreditech

Kreditech gained its popularity as one of the first start-ups to rely on social media data for issuing consumer loans, claiming to be able to do without traditional credit references altogether (Deville, 2013). Instead, major attention is given to alternative data about the credit requester, such as the information available on social media. Customers apply for a credit on one of the Kreditech's country-specific websites together with giving consent to access their profiles on websites such as Facebook, LinkedIn, eBay, etc. (Friedrich, 2018). Subsequently, algorithms assess the shared

data by calculating the likelihood of customers repaying the loan and provide the response in the mean of seconds (Huch, 2016, p. 69). In order to achieve the best response time in the industry, Kreditech heavily reduced the complexity of the credit analysis process, and simplified the traditional risk management (Huch, 2016, p. 69). Nevertheless, 20.000 observations or data points are stated to cover in total 8.000 variables in the process of approving the loan application (Huch, 2016, p. 69). The declared goal is to provide highly tailored financial services (consumer loans, credits) to those excluded from the traditional access due to, e.g., having not enough of the conventional data for credit reporting (Friedrich, 2018). At the same time, Kreditech specializes on short-term microloans (Huch, 2016, p. 69).

The in-house developed algorithms strongly rely on machine learning and Big Data analytics, constantly improving their predictive power with the growing number of customers: only in the first three years already three million applications were processed (Friedrich, 2018). The so-called digital footprint of users constitutes the core of Kreditech and is essentially important for its success. The activities and the information available on Facebook, Twitter, purchase history on Amazon and eBay, the apps installed on the user's devices, browsing behaviour, and even the precise data on how customers move around on the website, the way they fill the application form, etc. are all the examples of the major data of interest for creditworthiness assessment by Kreditech (Deville, 2013) (Friedrich, 2018) (Huch, 2016, p. 69). In order to receive a better sense of how extensive the evaluation actually is, (Deville, 2013) cites that loan applicant's "*public profile, friend list, email address, custom friends lists, messages, News Feed, birthday, chat status, work history, status updates, checkins, education history, groups, hometown, interests, current city, photos, website, personal description, likes*" from Facebook were required to be gained the access to. Furthermore, the information about the customer's friends on Facebook is also required to be accessed, in particular their "*birthdays, work histories, status updates, checkins, education histories, events, groups, hometowns, interests, current cities, photos, websites, personal descriptions and likes*" (Deville, 2013).

The data-centric approach of Kreditech inevitably requires strong acknowledgement of interfaces to the necessary sources of data. In addition to the existing abilities of a convenient access to social media data, the CEO of Kreditech also praises the rise of banking API's, contributing to this development by providing an API at Kreditech as well (Friedrich, 2018). The goal is to offer e.g., online retailers to add Kreditech's service to their operation, thus increasing the number of potential customers (Friedrich, 2018).

Little is directly known so far about the acceptance particularly of Kreditech, although the large number of its customers underscores the high popularity of Kreditech. The possibility to receive loan even with no relevant credit history, without the necessity to provide proof of income and of current debt (level) seems to be clearly an attractive offer for some (Huch, 2016, p. 69). The CEO of Kreditech names the growing ability of consumers to handle their own data as one of the reasons for being ready to share personal information for a good value proposition (Friedrich, 2018). Among others, (Huch, 2016) brings to the fore the Kreditech's cross-selling approach of additionally offering suitable financial products for particular customers, determined based on their shared data.

3.1.3 Social Credit System

The development of a modern credit system began in China relatively late. Meanwhile, the credit reporting progressed in China significantly, with current culmination in the form of the Social Credit System. The widespread of the Internet together with major inter-connectivity among the population and various businesses and shift of commercial activities largely into the Internet are often used to justify the eligibility of making use of all sorts of data from the Internet for the credit analysis purposes. Nevertheless, the decision to take into account social activities for credit analysis also unavoidably causes some critical reactions. (Huang, Lei, & Shen, 2016, p. 300) underscores the ability to infer the behaviour, personality, and economic status of individuals to assess their future affordability based on their online data from social platforms and online interactions.

In order to develop a more comprehensive understanding about the SCS, it is crucial to study the reasons behind this initiative, implementation specifics, and already conducted or still currently undergoing pilot projects.

Reasons behind the SCS

As already the name of the SCS reveals, the Chinese authorities are aiming to go simultaneously into social and financial directions. Hence, although the domain of the SCS comprises various areas, the central goal is to cover different social aspects in addition to the original financial matters. On the one side, the interconnection between social and financial aspects is of the main interest for the present evaluation, whereas the general impact of having a proper credit system on the economy of a particular country is mentioned only as a side remark. On the other side,

particularly individual credit is of the main importance, leaving aside the issues regarding enterprise credit and intermediary agencies.

Officially, the SCS is broadly underlined as being "*an important method to perfect the Socialist market economy system, accelerating and innovating social governance, and it has an important significance for strengthening the sincerity consciousness of the members of society, forging a desirable credit environment, raising the overall competitiveness of the country and stimulating the development of society and the progress of civilization*" (Creemers, 2015). This stated vision is summarized by (Chen & Cheung, 2017) into the following categories: increasing market efficiency, improving social governance, and building a harmonious socialist society.

The financial aspects are straightforward and rather obvious. A proper credit system is absolutely necessary for each modern economy. Taking into account the recentness of the organized credit reporting in China, the argument of the necessity to develop a credit system from the financial reasons is hence even easier to justify. Nevertheless, the question remains widely understudied of whether social behaviour delivers indeed proper insights to derive financial (un-)trustworthiness from it, and if so, how to properly translate it to the advantage of society (retaining wide acceptance and positive reception among the population). Consequently, a possible negative impact of mixing financial matters with social behaviour cannot be completely excluded. Some individuals would definitely be willing to try to game the system instead of conforming to the SCS rules, with for now unpredictable implications for the other individuals, and the official responses on such attempts.

The incorporation of the social aspect as an essential component in the credit analysis is what makes the SCS controversial and inconvenient. The key characteristics, which are declared to be developed in the social sense, include enhancing trustworthiness, developing moral, sincerity, ensuring integrity with socialist core values (Creemers, 2015). In order to achieve this, a comprehensive system of benefits and a broad sanction system are to be developed, i.e., to award a compliant behaviour and to punish nonconformity (undesirable behaviour). Due to the constantly growing number of the Internet users and, hence, more social interactions taking place online, linking data from e-commerce and social media activities into the SCS is the inseparable part of the declared social governance (Creemers, 2015), unavoidably leading to major discussions in this area. Again, there is obviously a chance for both false positives and false negatives, i.e., respectively awarding high social credit scores to those doesn't deserving it, and vice versa giving low social credit scores to trustworthy individuals.

Implementation specifics

The State Council of China stated for the SCS to come in 2020 into effect, i.e., assigning and managing a credit score for every citizen and each business entity that operates in the country (Creemers, 2015). Planning, construction, and all other preparatory steps are scheduled for the period between 2014 and 2020, involving "*all provincial, autonomous region and municipal People's Governments, all State Council ministries and commissions, all directly subordinate departments*" (Creemers, 2015) in addition to commercial organizations, who were allowed to conduct pilot projects in their business (operation) areas.

A remarkable characteristic of the SCS is the source of data to be included into calculating the credit score. On the one hand, rather traditional for credit reporting data, such as financial standing or criminal record (Kshetri, 2016, p. 302), is declared to be integrated into the SCS. On the other hand, nonfinancial (in particular widely defined as social) data is also aimed to be collected and processed. Hence, data from the various governmental departments and other official organizations is meant to be merged with data supplied by businesses (noteworthy that their credit scores are in turn also being computed).

Of a particular interest for the present work is the construction of credit system in the domain of Internet activities. First of all, an important requirement is to "*progressively implement the online real-name system*" (Creemers, 2015), which would provide means for the exact assignment of each online activity to a specific individual. Next, the requirement to "*vigorously move forward with the establishment of exchange and sharing mechanisms for online credit information and corresponding credit information in other areas*" (Creemers, 2015) would permit an extensive collection of any kind of online data, which is necessary to "*evaluate (...) the online behaviour of netizens*" (Creemers, 2015).

The logical follow-up concern is the way to receive high or low social credit score, together with respectively possible benefits or imposing sanctions. Naturally, the information impacting (and also being impacted by) scores of individuals include data from financial sector, data on taxes, etc. Additionally, e.g., data on public security is also taken into account (Cheng & Ou, 2014, p. 170). On the other hand, probably one of the most controversial attributes impacting social credit score is reported to be the information on political views and expressed opinions (e.g., on social media) regarding politically relevant issues (Kshetri, 2016, p. 302). In the official plan, it is expected that Chinese citizens behave patriotically (Creemers, 2015), whereas (Kshetri, 2016) concludes the overall requirement of not questioning or challenging the official viewpoints in order not to get own credit rating degraded. Moreover, "*information included in the rating may*

also include what books people read" (Kshetri, 2016, p. 302). In between, e.g., the performance at work also influences social credit score, with especially precisely assessing employees in education, health, and judiciary sectors (Creemers, 2015).

Potential sanctions in case of low score are also similarly far-reaching. (Diab, 2017) states that social credit score is expected to be used to handle the access to particular structural privileges. (Chen & Cheung, 2017) name denying to buy airline tickets or to travel on high-speed trains as other possible sanctions for a low score. The eligibility to occupy high-status or influential positions in public organizations and even private companies might be permitted only to holders of high scores (Kshetri, 2016, p. 302). Perhaps easier to justify are financial sanctions including denial of credits, e.g., to start a company, for housing, etc. (Kshetri, 2016, p. 302). (Chen & Cheung, 2017) also identify the possibility of restricting the access to education (e.g., private or elite schools) of children of individuals with a low score. Taking into account that social credit scores are meant to be publicly available, there is also an obvious implication on overall reputation of individuals (Chen & Cheung, 2017). Social network sites precisely support taking into account also friends and acquaintances by means of available connections on respective web resources, establishing particular interdependencies between social credit scores of different individuals (Kshetri, 2016, p. 302).

Pilot projects

Private credit service providers started to emerge in China only in the recent years. People's Bank of China released namely in 2013 the "Credit Industry Management Regulations", by which it allowed businesses to start offering credit services to consumers (Zhang, Xiong, Ni, & Li, 2015, p. 3). Since that time, dozens of such private credit services emerged. Another important milestone to "*the marketization of China's personal credit reporting industry*" (Huang, Lei, & Shen, 2016, p. 298) was reached in 2015, when People's Bank of China selected 8 private companies to develop and to implement Internet-based credit rating and ranking systems as a preparation for the SCS. The analysis of those particular pilot projects provides important insights into how the SCS's own implementation should look like for when it starts operating for each citizen in China. (Huang, Lei, & Shen, 2016) provide the categorization of the developed products into those offered by Internet and financial giants (Sesame Credit, Tencent Credit, Qianhai Zhengxin), by traditional credit rating agencies (Pengyuan Credit, China Credit Co., IntelliCredit), and by other private companies (Koala Credit, Sinoway Credit).

Sesame Credit, also known as Zhima Credit, is of a particular interest for this work due to its main source of information for computing the respective credit score, namely various online data, specifically from the e-commerce realm. Sesame Credit was developed by the Ant Financial Services Group, an affiliate of the Alibaba Group, making strong use of the huge amount of available data. Alipay, the Alibaba's mobile and online payment platform, currently counts 520 million registered users (Alipay, 2018). There have been also almost the same number of online buyers consolidated across Alibaba's different online shopping enterprises, such as Taobao and Tmall, in 2017 (Alibaba: cumulative active online buyers, 2018). This major advantage of having an exceptional access to millions of transactions in addition to large amounts of users-specific data is a strong base for evaluating financial repayment willingness and ability of consumers in the form of Sesame score. As the result, the creditworthiness of the borrowers is stated to be assessable more accurately, together with offering fine-grained credit-related services (Kshetri, 2016, p. 301). The lowest possible credit score is currently set to 350, while the highest achievable score is 950 points, with a person considered to be creditworthy in case of having Sesame score above 600 (Chen & Cheung, 2017, p. 9). The consumers' attributes taken into count include their identity features, behavioural preferences, credit history and performance capacity, generalized to spending habits, also closely evaluating what the money in question are going towards (Chen & Cheung, 2017, p. 9) (Tao & Zhang, 2016, p. 8). Remarkable is also the consideration of interpersonal relationships, i.e., lending and spending habits of users' connections (Chen & Cheung, 2017, p. 9). Many Chinese citizens are already experiencing the effects of their Sesame score, which affects, under certain conditions, "*the level of screening they are subjected to at airport security, the insurance premium they have to pay, their chances of adopting a pet from an animal shelter and even their placement on online dating services*" (Chen & Cheung, 2017, p. 9). At the same time, (Chen & Cheung, 2017) underline lack of clear and undisputed confirmations (evidence) of the ability of the Sesame Credit scoring system to accurately predict credit default.

Tencent Credit is affiliated with another Chinese Internet giant, namely Tencent, which operates predominately in the social media and online games area, recently also largely expanding into the field of mobile payment services. Tencent's social media mobile app WeChat counts currently ca. 980 million active users and the instant messaging platform Tencent QQ has almost 850 million active users (Most famous social network sites worldwide, 2017). Hence, similarly to the Sesame Credit, also the Tencent Credit is built on the top of a massive amount of online data from the social media and e-commerce. In the financial sense, Tencent consistently encourages its users to link their bank cards with the WeChat Pay service, offering certain payment convenience and rewarding them with specific deals (Kshetri, 2016, p. 301). More than 100 million WeChat and QQ users linked their traditional payment options with Tencent's payment system already by the

end of 2014 (Kshetri, 2016, p. 301). Tencent obviously aims to get the access to valuable transaction data, making credits available to its large userbase, who would be potentially disclosed from the possibility of getting credits otherwise. The Tencent Credit score supports the assessment of the creditworthiness of individuals under the consideration of their online shopping behaviour, activities on social networks, data from online games, etc. (Kshetri, 2016, p. 301). The creditworthiness of 50 million customers has been rated by Tencent already by 2015 (Chen & Cheung, 2017, p. 9). Moreover, the close cooperation of Tencent with banks and other companies in the financial sector resulted in founding in 2015 the WeBank, the first privately-held and the first online-only bank in China (Huang, Lei, & Shen, 2016, p. 298) (Lu, 2016, p. 592). Hence, the main target group of WeBank are naturally Internet users, which are in addition either not eligible for traditional bank loans (Lu, 2016, p. 593). In order to apply for a loan at WeBank the user takes his picture using a smartphone camera and submits the application online. The identity verification is conducted using a special face recognition system in accordance with the data provided by the Chinese Ministry of Public Security (Lu, 2016, p. 597).

Summing up, the convenience of the services facilitated by the introduction of Sesame Credit and Tencent Credit is obviously highly welcoming and beneficial for many consumers. At the same time, some others just cannot afford to stay outside of such online credit scoring systems, regardless of the potential sanctions and exclusion. The implications on privacy and individual rights are also an important concern required to be carefully taken into account.

The credit products of the rest six other pilot projects selected as the preparation for the SCS, described by (Huang, Lei, & Shen, 2016) in details, went the way from providing traditional financial services to extending their portfolio by additionally offering Internet credit scores or credit ratings. Qianhai Zhengxin is developed by PINGAN, a financial giant specializing in risk management and offering among others Internet investment and financing services. Pengyuan Credit makes use of public data from their Pengyuan credit reporting system, similar to China Credit Co. and IntelliCredit, which are also utilizing data available from their experience as traditional credit rating agencies. Koala Credit was launched as the result of a strategic cooperation between Lakala Credit Management Co. with China UnionPay and hundreds of other financial institutions, whereas Sinoway Credit was founded by four large financial companies to start providing an Internet-based credit reporting system.

Besides those 8 country-wide pilot projects launched as the preparation to the SCS, there are also multiple similar experiments running on the provincial level, targeting more limited but specific areas. Already in 2010, a programme with a close look on behaviour of citizens was started in

Suining County of Jiangsu Province. For bad behaviour, such as traffic violations or illegally petitioning higher authorities, points were deducted, contrary to giving points for good behaviour (Chen & Cheung, 2017, p. 6). Possible awards included faster promotions at work or shorter processing time of public housing applications (Chen & Cheung, 2017, p. 6). In Chongqing municipality, the authorities decided to create the so-called Red-Black-List system with public access to it through a specific government website. The idea is to place individuals with a high credit score on the Red List and individuals with a low credit score on the Black List respectively (Shan, 2017, p. 76). Hence, the Red List is thought to contain citizens to be presented as role models in regard to the correct behaviour for others, whereas the Black List should contain those who seriously breached the law or regulations (Shan, 2017, p. 76). Probably the most comprehensive programme of this kind was started by the Shanghai municipal government in 2016. The catalogue of items influencing the credit score contains several thousand items relevant for the business entities or concerning individual citizens. In order to get the idea of the dimension of covered aspects, the credit score is, e.g., influenced by the frequency with which a particular citizen visits his parents and whether they have enough food (Chen & Cheung, 2017, p. 6). The app called “Honest Shanghai” logically complements this project. It is stated that data to compute the credit score is collected in total from about 100 different government agencies in addition to data from industry associations, private companies and social media (Shan, 2017, p. 76). Some of the potential rewards include discounting transportation tickets, in contrast to the individuals with a low score struggling to get at all seats on, e.g., trains or planes (Shan, 2017, p. 76). Another country-wide example and an important milestone in the stepwise introduction of the SCS is the launch of the Credit China website, which already initially exposed for public access more than 1 million pieces of information on credit histories of some citizens and firms, predominately those involved in tax avoidance, who failed to follow court rulings, etc. (Kshetri, 2016, p. 302).

3.2 Social media for credit scoring under the GDPR

Mainly appliable law to regulate social media user profiling in credit scoring towards individuals in Austria is the data protection act (DSG) (Datenschutzgesetz, 2020) and the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679, 2020). Since DSG supplements GDPR (Relevant Data Protection Laws, 2020), and it is reasonable to expect that EU’s GDPR received much wider coverage in the academic publications, the decision is made to further concentrate on clauses in GDPR, additionally assessing DSG extensions, whenever applicable.

The GDPR, adopted on the 14th of April 2016, became enforceable on the 25th of May 2018. The overall aim is to respond to the recent technological advances in the data (handling) domain in order to protect all EU citizens from privacy and data breaches. The GDPR applies "*in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not*" (GDPR, Art. 3(1)) and "*where the processing activities are related to: (a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or (b) the monitoring of their behaviour as far as their behaviour takes place within the Union*" (GDPR, Art. 3(2)).

One of the central terms in the GDPR is the widened definition of personal data first as "*any information relating to an identified or identifiable natural person ('data subject')*" (GDPR, Art. 4). Since a natural person is identifiable under the account of "*all the means reasonably likely to be used*" (GDPR, Recital 26), not only single pieces of data, but also multiple data points that can be combined to create a record are defined to belong to the definition of personal data. Furthermore, "*natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers*" (GDPR, Recital 30). Finally, sensitive personal data is such that belongs to special categories of personal data as "*racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, (...) genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation*" (GDPR, Art. 9(1)). Remarkably, the GDPR is not applicable to the anonymized information, i.e., "*information which does not relate to an identified or identifiable natural person*", as such data is not seen as personal data anymore (GDPR, Rec. 26).

There are no indications that under the GDPR it would generally be prohibited to consider social media for credit scoring. At the same time, particular clauses are clearly applicable to a lesser or greater extent for such undertaking, addressed in the following subsections 3.2.1, 3.2.2, and 3.2.3 by respectively covering processing principles, rights of individuals, and data protection impact assessment, as regulated by the GDPR.

3.2.1 Processing principles

The GDPR contains extensive regulation regarding processing personal data, introducing certain novel concepts and improving requirements to existing approaches in the domain of privacy sensitive data processing. The main principles, to which processing personal data should adhere, are (GDPR, Art. 5(1)):

- Lawfulness, fairness and transparency: legal bases, fair reasons, and should be transparently justifiable.
- Purpose limitation: processing in a way that is strictly compatible with the declared purposes.
- Data minimization: not utilizing more personal data than what absolutely required.
- Accuracy: taking every reasonable step to ensure accurateness of used personal data.
- Storage limitation: obligation to timely erasing personal data that is not needed anymore.
- Integrity and confidentiality: measures to guarantee that personal data has not been improperly modified with security measures protecting personal data.

Furthermore, additional conditions apply to processing data belonging to special categories of personal data (GDPR, Art. 9).

A very specific point is the requirements on consent by data subjects. Naturally, the data controller should be able "*to demonstrate that the data subject has consented to processing of his or her personal data*" (GDPR, Art. 7(1)). At the same time, the request for consent should be shown to the data subject "*clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language*" (GDPR, Art. 7(2)). Another strengthening measure is the requirement to be "*as easy to withdraw as to give consent*" (GDPR, Art. 7(3)).

The newly defined concept of pseudonymization is a special case of processing personal data, namely "*in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;*" (GDPR, Art. 4). The usage of pseudonymization is prescribed to "*reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations*" and is "*not intended to preclude any other measures of data protection*" (GDPR, Rec. 28).

Not rarely the data needs to be transferred cross-border or just shared with 3rd parties during the processing phase. Personal data is permitted to be transferred to countries outside the EU only if the conditions laid down in the GDPR are met (GDPR, Art. 44). In case of the need to share personal data between data controller and data processor, it is important to keep in mind the requirement of being able to revoke the access to these data respectively if data subject files a corresponding request.

The novel introduction of certain responsibilities of data processor, in addition to those of data controller, is another major aspect of the GDPR. By the definition, controller "*determines the purposes and means of the processing of personal data*" during processor "*processes personal data on behalf of the controller*" (GDPR, Art. 4), although those might be also the same entity. Since personal data may be in use by both data controller and data processor, hence also the responsibility for each of them is regulated by the GDPR.

First of all, there are extensive rules and norms in place for how data processor is to be appointed by data controller, which requirements should be met, etc. (GDPR, Art. 28). Processors are obligated to maintain adequate documentation (GDPR, Art. 30), to cooperate with national supervisory authorities (GDPR, Art. 31), to comply with appropriate security standards (GDPR, Art. 32) and rules on international data transfers (GDPR, Art. 44-50). Consequently, processors may face private claims by the individuals for compensations (GDPR, Art. 79) and are liable to potential sanctions (GDPR, Art. 83).

In addition to principles generally regarding processing personal data, the GDPR also distinguishes in particular profiling of consumers, defining it as "*any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning (...) economic situation, health, personal preferences, interests, reliability, behaviour, location or movements*" (GDPR, Art. 4). Further important element of profiling is specified in terms of its intention regarding data subject, namely "*to take decisions concerning her or him*" (GDPR, Rec. 24).

Consequently, data subjects are also entitled to particular rights regarding profiling. First of all, data subjects have the right to avoid being "*subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her*" (GDPR, Art. 22(1)). Online credit application is even explicitly named as one of such examples (GDPR, Rec. 71). Provided that data subject gave explicit consent, there should nevertheless be the possibility to contest the decision (GDPR, Art. 22(3)). Importantly, data subjects have particularly the right to receive "*meaningful information about the logic involved, as well as the significance and the envisaged consequences (...) for the data subject*" (GDPR, Art. 13(2)).

At the same time, data controller is obliged to "*use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures (...) that factors*

which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject" (GDPR, Rec. 71), underlining specifically the demand to prevent "*discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect*" (GDPR, Rec. 71).

3.2.2 Rights of individuals

The rights of individuals are naturally not limited to the domain of profiling only. The GDPR extends and enhances the rights already enjoyed by the individuals under the previous legislation, introducing also some completely new rights, with the most important among them as follows:

- The right to erasure. Individuals become the right to be forgotten, i.e., to request the data controller to erase their personal data (in case certain conditions are met), and to inform third parties about changes, if applicable (GDPR, Art. 17).
- The right to restriction of processing. Individuals receive the right to restrict processing of their personal data under specified circumstances (GDPR, Art. 18). The reasons and other specifics are covered and similar to those in the subsection on profiling.
- The right to data portability. The data subjects should be able to receive their personal data "*in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance*" (GDPR, Art. 20).

The other rights directly regulated by the GDPR are: the right of access (GDPR, Art. 15), the right to rectification (GDPR, Art. 16), data breach notification (GDPR, Art. 34).

3.2.3 Data protection impact assessment

Further important aspect in the context of the GDPR is related to the potential data protection risks, their assessment and management. Recital 75 defines "*the risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage, in particular: where the processing may give rise to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented*

from exercising control over their personal data; where personal data are processed which reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures; where personal aspects are evaluated, in particular analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles; where personal data of vulnerable natural persons, in particular of children, are processed; or where processing involves a large amount of personal data and affects a large number of data subjects" (GDPR, Rec. 75).

Art. 32 GDPR outlines the risk-sensitive requirement to the implementation of organizational and technical measures in order to achieve a high degree of security in data processing. At the same time, the comprehensive Data Protection Impact Assessment (DPIA) is regulated in Art. 35 GDPR, i.e., the precise obligation for the documentation of conducted risk analysis and, if necessary, respective corrective measures. Finally, the necessary additions to the DPIA are indicated in the case of a need for consultation with the supervisory authority in Art. 36 GDPR. The aforementioned articles of the GDPR together constitute an important component in terms of data protection risks in the overall concept for data protection compliant data processing.

From the risk management point of view, it is an essential requirement for both the controller and the processor to adhere to the appropriate security measures: "*Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk*" (GDPR, Art. 32(1)). A set of specific measures is also listed for this purpose, including "*a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing*" (GDPR, Art. 32(1)(d)).

With regard to the data protection impact assessment, it must first be clarified whether the DPIA should at all take place. Art. 35(3) GDPR specifies when the DPIA is required, such as in case of "*a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person*" (GDPR, Art. 35(3)(a)) or "*a systematic monitoring of a publicly accessible area on a large scale*" (GDPR, Art. 35(3)(c)) taking place. The supervisory authority also prepares and

publishes lists of kinds of processing operations for which the DPIA is mandatory (GDPR, Art. 35(4)) and for which the DPIA is not required (GDPR, Art. 35(5)). The following parts are the essential elements of the DPIA:

- "*a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller*" (GDPR, Art. 35(7)(a));
- "*an assessment of the necessity and proportionality of the processing operations in relation to the purposes*" (GDPR, Art. 35(7)(b));
- "*an assessment of the risks to the rights and freedoms of data subjects referred to in paragraph 1*" (GDPR, Art. 35(7)(c));
- "*the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned*" (GDPR, Art. 35(7)(d)).

3.3 Ethical issues of social media data in credit scoring

The relation of the provided research to a rather controversial undertaking to use social media data in credit scoring requires to incorporate a discussion on ethical issues, i.e., issues concerning "*moral correctness of specified conduct*" (Definition of Ethics by Oxford Dictionary, 2021). In the context of the present work, approach to assess the ethical implications of data science elaborated by (Floridi & Taddeo, 2016) perfectly suits to be a foundation to build the discussion in the following subsections upon.

Initially, the wide-ranging works of Norbert Wiener, a professor of mathematics and engineering at MIT, in the middle of the 20th century led to the creation of a new branch of ethics coined as information ethics, capable to be effectively applied to identify, analyse, and resolve ethical issues associated with all kinds of information technology, such as computers, computer networks, radio, television, telephones, news media, journalism, even books and libraries (Bynum, 2015). Later, the observations of the professor Walter Maner on the significant impact of specifically computer technology on ethical questions led to the establishment of the separate field of computer ethics to respectively study ethical issues in conjunction with computer technology (Bynum, 2015). Taking the developments of the information and computer ethics as basis, (Floridi & Taddeo, 2016) propose a data-centric level of abstraction to cope specifically with the ethical impact of

data science. To cover the different aspects, the approach by (Floridi & Taddeo, 2016) suggests to evaluate ethical issues respectively related to data, algorithms, and corresponding practices as follows:

- ethics of data covers study and evaluation of moral problems related to generation, recording, curation, processing, dissemination, sharing, and use of data;
- ethics of algorithms covers study and evaluation of moral problems related to artificial intelligence, artificial agents, machine learning, and robots;
- ethics of practices covers study and evaluation of moral problems related to responsible innovation, programming, hacking, and professional codes.

The following subsections 3.3.1, 3.3.2, and 3.3.3 provide the in-depth discussion of the potential ethical issues respectively posed by usage of social media data in credit scoring on these three directions.

3.3.1 Ethics of data

Prior to the discussion on ethics of data, it is important to align on the technical considerations regarding data in the context of social media data usage in credit scoring, covered by the following short outline.

First of all, the required data has to be collected from social network sites, such as Facebook, Instagram, Twitter, LinkedIn, etc. In addition to service-specific user profile and user-generated content of the person under credit scoring, also network of connections could be collected. At the same time, meta-data or even in-depth analysis of single entities in this network of connections can similarly be required, raising the need to collect also respective data on the users, groups, and specific points of interest, to which the connections exist. Finally, it might be of a certain interest to track the potential debtor for some period of time, instead of retrieving only a snapshot of the current social media data.

From the technical point of view, the data can be either scrapped by web crawling agents or more directly accessed by the means of a service-specific API. Some of the social network sites are actively taking actions to prevent crawling data from their services, while APIs are widely established to provide the required access, and, hence, might be preferred. The wide application possibilities of social media data nowadays also triggered the rise of various supportive tools to even further ease the data collection process. Then, particular data is usually required to be persisted, be it in form of flat files, in document-based databases, in graph databases, or in more

traditional relational databases. An important possibility to consider is the nowadays' widespread usage of cloud services, i.e., instead of storing data on-site to keep it in external storage facilities managed by third parties. The type of data to be persisted ranges from data merely required for the offered services to data prescribed to be archived in order to satisfy particular regulatory requirements. A possible overall solution to satisfy the data storing challenges is to conduct a separation into the so-called hot and cold data, or more specifically to make use of, e.g., the following three tiers: the in-memory tier for data processing, the on-disk tier for intermediate outcomes, and the cold-data tier (i.e., an offline backup) for the archive database. At the same time, stored data should also be accessible by internal processing routines same as by third parties, just differencing on what data by whom. The most obvious choices to enable data access is by the means of providing some webservice or an API with particular access rights.

The focus of the ethics of data on moral problems related to generation, recording, curation, processing, dissemination, sharing, and use of data results in risks of identifying types of individuals and up to the exact re-identification of individuals as the major ethical concerns related to data (Floridi & Taddeo, 2016, p. 3).

Identification of types of individuals

The possibility to identify individuals by their specific characteristics, such as individuals' age, ethnicity, gender, etc., may lead to serious ethical problems in the context of using social media data in credit scoring. Here, discrimination is a major ethical issue that could arise, i.e., in the process of credit scoring unethically differentiating individuals upon the identified affiliation to a specific group of people. In the recent research in this field of study such ethical threats are described as breaching group privacy (Floridi & Taddeo, 2016, p. 3). For instance, different credit opportunities based on age, ethnicity, or gender are respectively the examples of ageism, ethnicism, or sexism.

From the implementation point of view, collecting more data than required is an important facilitator for the unethical identification of groups of individuals to take place, i.e., creating high potential for group privacy breach. Hence, it must be precisely justified what social media data is to be collected for credit scoring, and the overall issue of data choice to be handled in accordance with the aim of the further data processing, i.e., collecting data only truly needed, is often a legal requirement (e.g., data minimisation principle by the EU's GDPR (Regulation (EU) 2016/679, 2020)). Furthermore, as (Backer, 2017, p. 6) underscores, not everything what can be collected could indeed contribute to credit scoring, at the same time leading to potential social, economic,

or political threats by unnecessarily collecting sensitive or simply irrelevant data. (Toch, Wang, & Cranor, 2012) exemplified in particular the potential negative consequences of collecting location-based social media data, even if justifiable for personalization as the main aim of the further data processing. (Backer, 2017, p. 6) pointed to the interchange between data collection and data processing as follows: the choice of data (e.g., social media data) fundamentally affects the information extractable from this data (e.g., credit score), whereupon the assessment of required information constitutes the scope of the data to collect.

Similar to the considerations for data collection, storing more data than required is also capable to facilitate the unethical identification of types of individuals. Thus, it should be carefully taken care for what, e.g., social media data truly needs to be stored and which data can instead be processed on-the-fly, e.g., to compute credit score. Directly related to this decision is the potential negative effect of (sensitive) data leakage as the result of some security breach. Next important influencer on potential threats by data processing and access is the decision for what period, and how is data retained (Backer, 2017, p. 6). Appropriate security measures should be in place to ensure sufficient protection of stored data, otherwise potentially unethically harming data confidentiality, integrity, etc. through an unauthorized access (di Vimercati, Foresti, & Samarati, 2012). Social media data collected and processed for credit scoring purposes, if stolen, would increase in first place threats of online phishing, social engineering, online identity theft (Al-Daraiseh, Al-Joudi, Al-Gahtani, & Al-Qahtani, 2014, pp. 132-133) (Gao, Hu, Huang, Wang, & Chen, 2011, p. 59). (di Vimercati, Foresti, & Samarati, 2012) describe in detail the approaches and potential threats in terms of managing and accessing data especially in cases when apparently strong protection measures are in place.

Re-identification of individuals

Following the common practices or even legal requirements for processing personal data (such as those defined by the EU's GDPR (Regulation (EU) 2016/679, 2020)), either anonymization or at least pseudonymization is to be applied to reduce the risks to the individuals and facilitate service providers meeting their obligations (Regulation (EU) 2016/679, 2020). At the same time, the possibility to re-identify people, i.e., to identify the exact individuals from the anonymized data, may lead to serious ethical problems in the context of using social media data in credit scoring. Here, similar to the identification of specific characteristics of individuals, discrimination is also a major ethical issue that could arise, i.e., in the process of credit scoring unethically differentiating between re-identified individuals. For instance, different credit opportunities based on exact individual to whom they are offered is an example of such unethical misconduct. In

general, re-identification of individuals could lead to privacy breach and various forms of misuse, more on which in the data misuse subsection of ethics of practices.

From the implementation point of view, collecting more data than required and storing data for longer than needed are also important facilitators for the unethical re-identification of individuals, similar to the case of identifying specific characteristics of individuals. In this context, worth noting are, in particular, the modern capabilities of de-anonymization, i.e., the procedure of utilizing sophisticated data mining techniques to re-identify individuals in anonymous data sets (Ali, et al., 2018, p. 5). De-anonymizing attacks to re-identify particular social media users pose a major privacy threat (Gao, Hu, Huang, Wang, & Chen, 2011, p. 58). (Ali, et al., 2018, p. 5) cite various effective techniques for precise and robust de-anonymization attacks from social media data. In addition to or as the result of de-anonymization there are different possibilities of data processing for unauthorized usage scenarios, causing respective privacy or other harm to individuals. Some of the examples of potential unethical misuse of data collected for credit scoring in other scenarios include data processing for personalized marketing (advertisements), inadmissible data transfer, etc.

3.3.2 Ethics of algorithms

Prior to the discussion on ethics of algorithms, it is important to align on the role of algorithms for data processing in the context of social media data usage in credit scoring, covered by the following short outline.

The main aim of the data processing stage is to actually conduct credit scoring. For this to happen, it is first required to pre-process the collected data for the subsequent analysis. There are many different methods of how to approach this task, often making use of data ingestion and staging, data extraction, transformation, and loading, etc. Similar among all of them is the goal to prepare the data for the respective data analysis algorithms. As next, the techniques for text mining, image recognition, geolocation analytics, and graph data analysis are in place in order to construct a complete social profile of the potential debtor. Eventually, the outcomes should be made available for the credit score computation. A specific underlying model for credit scoring based on social media user profiling is another crucial component in the data processing considerations. The model should guide and actively support all stages of the process of creating social profiles, deriving appropriate conclusions, etc. Roughly speaking, the following phases would normally take place: analysis of past cases in conjunction with social media data to construct the model;

using the constructed model in the production to conduct credit scoring; maintaining the underlying model by the terms of continuous adjustment and improvements.

The focus of the ethics of algorithms on moral problems related to artificial intelligence, artificial agents, machine learning, and robots results in epistemic concerns (e.g., inconclusive, inscrutable, or misguided evidence), normative concerns (e.g., unfair outcomes or transformative effect), and traceability concerns as the major ethical concerns related to algorithms (Floridi & Taddeo, 2016, p. 3) (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016).

Epistemic concerns

The application of algorithms to derive conclusions from data used as or processed to produce evidence, may lead to serious ethical problems in the context of using social media data in credit scoring. Produced knowledge (hence, epistemic concerns) might be inconclusive, inscrutable, or misguided (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, pp. 4-5), potentially leading to faulty results of conducted credit scoring.

There are different reasons or situations for the potential occurrence of inconclusive or ambiguous evidence. The conclusions derived from data using some sort of machine learning or inferential statistics approaches although produce probable yet inevitably uncertain knowledge, with additional procedures applied to quantify this uncertainty (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 4). Reversely, there is always (even if very small) probability for the derived knowledge to be faulty, as in case of conducting credit scoring from social media data to wrongfully determine the credit score of the credit applicant. Incorporating identified correlations in data for the purpose of credit scoring, although these are rarely sufficient to prove causal connections (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 4), is another possibility to produce inconclusive evidence. In other words, the existence of certain correlations between individuals' social media data does not prove their similarity in regard to their credit score. Furthermore, the legitimacy and coherence of connections between raw data and its interpretations should be carefully assessed (Backer, 2017, p. 6). The information on connections of particular user is usually of a high value, although social media connections are often merely an agreed link between two users regardless of their offline relationship (Gao, Hu, Huang, Wang, & Chen, 2011, p. 57). Hence, resulting conclusions bear potential to be largely faulty. In this sense, the widely used process of collaborative filtering introduce, e.g., otherwise non-existing links between users based on their specific common characteristics, similarly to how credit scoring based on social media data would work. Nevertheless, the similarity between users in one

context is not necessarily transferable to other contexts, with an otherwise assumption posing ethical issue of potential misclassification of (i.e., wrong conclusions about) individuals (Toch, Wang, & Cranor, 2012).

There is naturally also the expectation (often a legal requirement) on the accessibility of the process to derive conclusions from data (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 4). In other words, the connection between data and produced evidence is expected to be intelligible, open to scrutiny and possibly even to critique (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 4), specifically in the context of using social media data in credit scoring. Incorporation of modern sophisticated algorithms with weak explainability capabilities in the process of deriving conclusions from data (as this could be the case of using social media data in credit scoring) transfers the focus of decision making to a more abstract operational level, further facilitating non-transparency and inscrutability of produced evidence (Backer, 2017, p. 6). The difficulty to provide accessible explanations and gaps in understanding the link between available data and based on it generated conclusions are major ethical concerns that negatively contributes to the produced evidence, leading to respective limitations (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 4).

Furthermore, recalling Shannon's mathematical theory of communication reminds that algorithms through processing data can never exceed the input (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, pp. 4-5). That means, the derived by algorithms conclusions can be only as reliable or as neutral as their underlying data, hence, bearing the potential to lead to unethical misguided evidence in the context of using social media data for credit scoring.

Normative concerns

As opposed to covering potential ethical issues arising during application of algorithms to derive new knowledge (i.e., epistemic concerns), normative ethical concerns are those resulting from the potential implications or the overall effects (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 5). An example of a potential normative ethical issues is an unfair outcome (with observer-dependently defined "fairness"), e.g., a discriminatory effect even as the result of conclusive, scrutable and, well-founded evidence (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 5).

Another major normative ethical issue is the potential transformative effect (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 5), e.g., a negative social impact as the result of using social media data in credit scoring. Since large portions of social interactions takes place nowadays

online, hence, credit scoring based on social media data would unavoidably have a strong social impact as well. The two of the most important and broadest transformative ethical issues regarding social interactions, which would possibly arise as the result of utilizing social media data for credit scoring, namely those of individual behaviour (i.e., changes in ones' self) and social exclusion (i.e., interpersonal changes), are exemplified next.

Naturally, many people tend to regret some of their posts on social media. (Wang, et al., 2011) conducted an in-depth research dedicated to study regrets on Facebook. Their won insights are clearly transferable to users' potential perception of their activities on social media in the context of using social media data for credit scoring, hence, also possible negative impact on their online social behaviour as the result of an unethical transformative effect. Posts that are most regretted are also those, which would negatively affect credit scores of respective individuals. (Wang, et al., 2011, pp. 4-6) identified in their study sensitive content, content with strong sentiment, and lies and secrets as the main categories of such posts. One might think of an option to just avoid posting content, which will be regretted later about. Nevertheless, there seems to always exist situations when such posts reach social media. Due to (Wang, et al., 2011, pp. 6-7) the most common reasons for that are accidents, intended purposes, unforeseen or ignored consequences, and unfamiliarity with or misunderstanding of social network sites. Eventually, most of the social media users develop particular strategies to avoid posts that they would later regret. Similarly, posts which would also negatively impact one's credit score would need to be appropriately handled or potentially cause harm. Some of the most widespread strategies to prevent self-harmful posts and accordingly adjust individual behaviour include rules for information sharing, delays, declining or ignoring connection requests, self-censoring, self-cleaning, reading but not posting, multiple accounts for auditing, fake names or status (Wang, et al., 2011, pp. 7-9). An intention to alter own online behaviour, knowing that personal data is watched by third parties (government, private companies, financial institutions), is also confirmed by research of (Zuiderveen, et al., 2018) and (Dinev, Hart, & Mullen, 2008). Suspicion that own activities are monitored at some point of time leads to adapted behaviour, often just trying to escape attention through producing fewer digital traces (Zuiderveen, et al., 2018, p. 87). Some other reasons for behaviour modification result from concerns of individuals to be oversimplified, taken out of context, having part of their identity mistaken for the whole of their identity (Dinev, Hart, & Mullen, 2008, p. 221), as that could happen in computing individual's credit score from social media data. Moreover, privacy concerns are identified as having tremendous effect on spontaneity, creativity, productivity, and other psychological effects (Dinev, Hart, & Mullen, 2008, p. 221).

On the interpersonal level, the most probable unethical consequence of using social media data for credit scoring is an increased social exclusion. The respective phenomenon of unfriending exists since the very first appearance and growing popularity of social networks. The most common types of friends that are unfriended, and the reasons for unfriending are in-depth studied by (Sibona, 2014), for now not directly studying an importance of gaining personal advantage from being friend with certain individuals or not. Nevertheless, the main findings regarding current unfriending behaviour are definitely of a high importance for assessing the potential for growing social exclusion on social media as the result of a more extensive usage of social media data, in particular for credit scoring. Unfriending is rather a common aspect on social media, with main categories of individuals that are most commonly unfriended being very diverse, ranging from high school fellow students and colleagues from work to family members, etc. (Sibona, 2014, p. 1680). At the same time, the most common reasons for unfriending are much less diverse, hence leading to a growing risk that same individuals would get unfriended by different users, i.e., worsening social exclusion for some. Posts on polarizing topics, such as on political or religious issues, are among the leading reasons for unfriending (Sibona, 2014, p. 1681). In other words, one of the main strategies behind unfriending behaviour is the lowest common denominator approach by avoiding controversial topics altogether (Sibona, 2014, p. 1677). In case social media data starts to be utilized for credit scoring, it is, hence, reasonable to assume that many social media users would similarly attempt to avoid potentially getting their credit score lowered as a consequence of being friend with some specific (categories of) users. This could also result in certain topics slowly disappearing from online social interactions. Under the assumption that factors influencing credit scores are known, this would create potential for the unethical social discrimination based on less favourable for a higher credit score personality traits, demographics, etc.

Traceability concerns

Last but not least are traceability concerns, i.e., particular difficulty to debug algorithmic activity in terms of tracing those personally responsible for the single stages and the complete process of applying the algorithms in particular use case (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 5). In the context of social media data usage in credit scoring, the potential inability to always identify those who should be held accountable for the possibly caused harm by the conducted credit scoring also poses a major ethical threat.

3.3.3 Ethics of practices

Finally, the discussion on ethics of practices covers the questions regarding liabilities and responsibilities of people and organizations in charge of all of the practices around social media data usage in credit scoring, such as those responsible for data processes, strategies, and policies (Floridi & Taddeo, 2016, p. 3).

The focus of the ethics of practice on moral problems related to responsible innovation, programming, hacking, and professional codes results in consent concerns, user privacy concerns, and misuse concerns as the major ethical concerns related to corresponding practices (Floridi & Taddeo, 2016, p. 3).

Consent concerns

Explicit consent is an ethical (and often legal) requirement, widely covered in the academic publications of social media research (Hunter, et al., 2018) (Moreno, Goniu, Moreno, & Diekema, 2013). Competence, being adequately informed, and voluntariness are some of the most important general ethical requirements for a valid consent (Komesaroff & Parker, 2009).

Competence to make a certain decision, e.g., to decide on own social media data usage in credit scoring, refers to the ability to understand and to believe the provided information (Komesaroff & Parker, 2009). Hence, being adequately informed is a crucial precondition to make a competent decision. The information usually provided includes, but is not limited to, explanations of concrete rights and obligations, the purpose and duration of data collection, storage, and processing, procedures to be undertaken, potential risks and benefits, etc. (Nijhawan, et al., 2013) In the context of using social media data for credit scoring, there are, thus, ethical issues to provide the required information adequately and to ensure the potential debtor is also capable to make competent decision of whether to participate. The necessity to explain in particular technically challenging process of collecting, storing, and processing social media data for conducting credit scoring additionally increases potential threats resulting from these ethical issues. Noteworthy is, e.g., the situation with nowadays widely used so-called behavioural profiling, i.e., creation of user profiles based on various seemingly unrelated online information and activities aggregated altogether (Toch, Wang, & Cranor, 2012). Since for most of the users it is difficult to clearly understand how behavioural profiling is done in detail, i.e., which activities are included, to what extent, and which impact do they have, the decision of whether to consent behavioural profiling is hence very challenging (Toch, Wang, & Cranor, 2012). An unforeseeable impact of behavioural profiling for the users thus results in unethical lack of transparency of the overall credit scoring

process to make a competent decision on whether to consent. Furthermore, (Komesaroff & Parker, 2009) reiterate critics of conventional definition of competence limited to cognitive ability and the need to recognize the impact of emotions, values, intuitions, personal, social, and other contextual factors as well.

Special attention deserves also the process by which the consent is to be given. The so-called "behavioural lock-in" modes of giving context, e.g., providing a manifestation of consent by clicking a specific text like "I agree", are considered by some researchers as insufficient, hence, as those potentially leading to ethical concerns (Hunter, et al., 2018, p. 345). To receive quick access to the service of interest, e.g., credit scoring based on social media data, the expectation on user to thoroughly read and to take enough time to understand the (mostly) lengthy terms and condition of service, prior to accepting them, is often faulty (Hunter, et al., 2018, p. 345).

Privacy concerns

The concept of privacy underwent in the past various definitions (starting with the right to be left alone) and nowadays comprises different meanings depending on the context (Trepte, 2020). For the ethical discussion here, privacy is the right of individuals to determine when, how and to what extent information about them is disclosed to others (Turculeț, 2014, p. 968) (Trepte, 2020, p. 2).

The centrality of privacy for the evaluation of ethical issues (Hagendorff, 2020, p. 102) results in privacy being already present to a larger extent as part of various ethical concerns: in the subsections 3.3.1 on ethics of data and 3.3.2 on ethics of algorithms the technical specifics of the possibilities for privacy breaches are outlined, with the subsection 0 on data misuse providing an extensive overview of the potential consequences after privacy breach took place. The respective ethical issues of using social media data in credit scoring regarding privacy that are covered in these subsections are not repeated here.

Particularly for the discussion on ethics of practices, it is important to also consider privacy concerns that arise from the insufficient practices leading to potential privacy breaches of individuals (Turculeț, 2014, p. 969). In the context of conducting credit scoring under the usage of social media data, there is an ethical responsibility of the credit scoring service provider to protect credit applicant's social media personal data, i.e., to ensure the appropriate privacy protection measures are in place. Thus, the possibilities for a potential privacy breach are kept limited. Another ethical concern is the requirement to prevent individual's personal information taken out of its context (Turculeț, 2014, p. 969).

Misuse concerns

One of the major sources of data misuse is through inappropriate usage of data as defined when it was initially collected, i.e., abusing granted access and legitimate permissions (Gafny, Shabtai, Rokach, & Elovici, 2010, p. 3), e.g., instead of using social media data (solely) for credit scoring, as it was presumed, but for other purposes. The wide variety of possible misuse cases in the context of utilizing social media data for credit scoring, as discussed below, places misuse issues as major ethical concerns overall. A privileged insider can, for example, exploit own legitimate ability to access sensitive data from private reasons or on a far larger scale (Shabtai, Bercovitch, Rokach, & Elovici, 2014, p. 1), e.g., those involved in the process of conducting credit scoring unethically accessing the underlying social media data. An employee of some financial institution accessing personal or specifically financial data (e.g., credit score) of one of its customers with whom having some sort of private relationship is also an unethical data misuse incident. Private information about a company's customers (e.g., those whose social media data used for credit scoring) can be sold, e.g., to a competitor. In the political domain, members of ruling parties might be interested in sensitive data (e.g., private social media data or the resulting credit score) of their political opponents or critics. An important side remarks is the fact that research and surveying on the dimensions of how often and at which scale data misuse actually takes place is specifically complicated since data misuse incidents are often not reported to prevent the loss of reputation (Shabtai, Bercovitch, Rokach, & Elovici, 2014, p. 2). In the following, two specific data misuse directions, namely by government or authorities and by service provider or intruders, are explicated.

The reasons behind data misuse conducted by authorities can be very diverse, ranging from unjust personal enrichment to discreditation of opponents, or even worse, with gained access to social media data of the underlying credit scoring potentially facilitating such unethical behaviour. Possible data misuse in conjunction with government surveillance leads to Internet users' worry of a possible privacy breach, as confirmed by extensive research on this topic conducted by (Dinev, Hart, & Mullen, 2008). Surveillance is commonly understood as a process of collecting and processing personal data for the purpose of managing or influencing those, whose data was gathered (Dinev, Hart, & Mullen, 2008, p. 214). Government online surveillance aims at timely detection and prevention of fraud and security breaches, terrorist activities, and other crimes through careful assessment of citizen online behaviour. The key to success of online surveillance constitutes comprehensive profiles of individuals based on vast amount of data available, i.e., the more qualitative data is possible to collect the better. At the same time, these activities result in

growing concern in population to thereby lose or at least experience some harm to their privacy (Dinev, Hart, & Mullen, 2008), e.g., unethically utilizing social media data used for credit scoring, potentially also the outcomes of the conducted credit scoring, for government surveillance. One of the main factors that enable government intrusion is often seen in information asymmetry, i.e., possession of more or better information by one party than by the other (Dinev, Hart, & Mullen, 2008, p. 227). The recent growth in information asymmetry is caused by increased technological threats, against which governments in many countries deploy comprehensive measures that enhance government authority to obtain personal information about citizens from private sector (Dinev, Hart, & Mullen, 2008, p. 227), among which also usage of social media data for credit scoring could become. (Dinev, Hart, & Mullen, 2008) found also no evidences that the population sees government surveillance measures as justified in the face of modern threats in the domain of Internet technologies. The potential access of authorities to phone records, web-based transactions, e-mails, voice mail, etc. led to growing fear of government intrusion (Dinev, Hart, & Mullen, 2008, p. 216). The so-called political microtargeting is another alarming trend in the context of potential data misuse in the political domain. Online political microtargeting aims to target narrow categories of voters with fine-tuned messages based on behavioural analysis of data gathered about individuals, such as demographic characteristics, lifestyle, etc. (Zuiderveen, et al., 2018, p. 83), which the evaluation of unethically misused social media data or conducted credit scoring would facilitate. Political parties are thereby capable to utilize vast amount of available online data to choose policy stances that best match to targeted voter, e.g., student benefits for students, or family aid for families (Zuiderveen, et al., 2018, p. 83). Personalized political communication, tailored to individual voters is clearly more effective, although at the same time inevitably bearing certain major ethical risks. One of the most obvious threats is the threat to privacy through gathering massive amounts of individuals' data to infer, in particular, political preferences among other sensitive data. Manipulations and political exclusion are some other major ethical threats resulting from political microtargeting (Zuiderveen, et al., 2018, pp. 87-88) that could misuse social media data utilized for credit scoring, when gained access to. On the one hand, political microtargeting makes effectively use of modern capabilities to identify individual voters which are more likely to be convinced. These particular voters are then targeted with specific information that match their interests and vulnerabilities to increase their engagement. At the same time, through same techniques it is similarly aimed to minimize voter engagement of individuals more favourable towards rival parties. Microtargeting gives a political party the possibility to present it as a different one-issue party to different individuals, i.e., highlighting one different issue for each voter, leading to biased perception regarding priorities of that party among voters, and lack of transparency about the party's promises. Here, political polarization and spread of misinformation are also commonly used in targeted information (Zuiderveen, et al., 2018, p.

87). Citizens are thus becoming objects of manipulation. On the other hand, a political party can easily not advertise to individuals who are anyway not expected to vote, or whose support is expected not to be necessary for a win in a certain area (Zuiderveen, et al., 2018, p. 88). As the result, certain voter groups are ignored, i.e., excluded from the major initial source of information on political issues. This leads in turn to an underrepresentation of certain groups in a democracy (Zuiderveen, et al., 2018, p. 88).

Data misuse conducted by credit scoring service providers or some malicious third party unavoidably possesses certain similarities to the data misuse scenarios in the political domain. One of the main differences lies in intention for unethical data misuse, with financial interests clearly coming to the fore. Service providers are also the first place where individual's social media data for credit scoring would arrive, only afterwards being possibly passed to third parties. Private companies are largely making use of the increased technical capabilities of information systems to collect, to store, and to process data required to derive knowledge on consumer preferences based on constructed comprehensive profiles, subsequently utilizing it for various commercial purposes. Thus, there is an ethical concern of social media data applied in credit scoring being misused by service provider. Loss of privacy for individuals is also in this case one of the major potential unethical outcomes of such developments. Submits an individual own personal information during the process of receiving certain services (such as allowing access to social media data for credit scoring), then many different participants are involved in the further flow of such data. In addition to justified parties (such as Internet service providers, financial institutions, advertisers) other parties might also easily be illegally involved with help of various spyware, keyloggers, hacking, etc. (Dinev, Hart, & Mullen, 2008, p. 217). Negative consequences for individuals usually occur also at a much later time, further deepening the problem of correct assessment how and why data misuse with possible privacy violation took place. Unobtrusive observation of Internet users is largely enabled by a constant build-up of vast amounts of different traces from each individuals' activity, leading to the generation of detailed digital footprints of user's preferences, interests, behaviour, etc. Since companies are unavoidably interested in stimulating their potential consumers to expose as much of such data, the responsibility is hence often directly attributed to private sector for making consumers vulnerable (Dinev, Hart, & Mullen, 2008, p. 215). A special case of loss of privacy through data (in particular, social media data) misuse is the so-called identity theft. (Al-Daraiseh, Al-Joudi, Al-Gahtani, & Al-Qahtani, 2014) conducted an extensive research of identity theft in social media. In case of studied population, identity thieves aimed primarily at acquiring full name, date of birth, hometown, school information, data on bank accounts, relationship status, hobbies, and interests of their victims (Al-Daraiseh, Al-Joudi, Al-Gahtani, & Al-Qahtani, 2014, p. 133), with all of such

information potentially available when using social media data in credit scoring. Main factors contributing to success of identity thievery include lack of knowledge on how to protect own online identity, which laws and regulations are in-place to support respective protection and what are specific privacy policies on particular social network sites, the overconfidence in social media providers, together with enormous growth of number of users, and those willing and capable to commit cybercrimes (Al-Daraiseh, Al-Joudi, Al-Gahtani, & Al-Qahtani, 2014, p. 133). Furthermore, (Al-Daraiseh, Al-Joudi, Al-Gahtani, & Al-Qahtani, 2014, pp. 132-133) provide the following extensive list of the different ways for online identity theft to happen: data breach, friendly fraud, computer hacking, dumpster diving, skimmers, stolen wallet, mail theft, shoulder surfing, account takeover, spam attack, malware, spyware, social engineering, online phishing, phone phishing, romantic fraud, spoofing, job posting. Eventually, stolen data is used in a variety of ways, with the most important among them are in particular (Al-Daraiseh, Al-Joudi, Al-Gahtani, & Al-Qahtani, 2014, p. 134): engaging in illegal activities, obtaining a cell phone account, illegal use of credit card accounts, obtaining bank loans, spending victim's checking and saving accounts, receiving a new ID, unauthorized access to utility accounts, black market sales. Several other researchers dedicated their works to identity theft: (Reznik, 2012) investigates in-depth mainly the two major categories of identity theft on the Internet, namely creation of a fictious profile of victim without the victim's permission, and gaining an unauthorized access to the victims' accounts by stealing their credentials; (Bilge, Strufe, Balzarotti, & Kirda, 2009) exemplarily conducted two different identity theft attacks to access user personal information on popular social network sites to study the easiness of actually executing an identity theft. Thus, multiple data misuse scenarios in the context of using social media data in credit scoring pose real ethical threats.

4 Taxonomy planning

The planning phase covers the defining aspects of the taxonomy, such as software engineering knowledge area that it is associated with, objectives and subject matter of the taxonomy, taxonomy structure and procedure types, and to identify sources of information for taxonomy development, as described in the following.

[A01] Define SE knowledge area. The software engineering knowledge area of the taxonomy is selected under the consideration of the knowledge areas defined in the Software Engineering Body of Knowledge (SWEBOK) (Bourque & Fairley, 2014). Evaluation of the explainability techniques requires thorough understanding of the algorithms in question, ability to assess these algorithms from the different aspects, often reasoning over their basic building blocks or components. Hence, among the 15 of the respectively defined knowledge areas, one that comes the closest to the topic of explainability techniques of social media user profiling approaches for credit scoring is the knowledge area "Computing Foundations", and more precisely its sub-area "Algorithms and Complexity".

[A02] Describe the objectives of the taxonomy. The main objective of the developed taxonomy is to provide a mapping scheme between the approaches of social media user profiling applicable to credit scoring on the one side and explainability techniques of these approaches on the other side. As an additional level, the components (both traditional and alternative) affecting or capable to affect credit scoring models are also included, completing the elaborated view on the social media user profiling approaches for credit scoring and techniques to achieve their explainability. The existing research lack studies on explainability techniques specifically for the social media user profiling techniques (applicable to credit scoring), as discussed in the chapter on the state-of-the-art. Hence, the proposed mapping scheme supports developers and decision makers during the process of choosing a specific user profiling approach under certain explainability requirements.

[A03] Describe the subject matter to be classified. Explainability techniques for social media user profiling approaches applicable to components affecting credit scoring models is the subject matter of the developed taxonomy.

[A04] Select classification structure type. For the proposed taxonomy, a polyhierarchical classification structure is selected, justified by the aim to allow complex relationships between single levels of the taxonomy (Harpring, 2010). Thereby, single categories of credit scoring model components can be easily associated with multiple social media user profiling approaches to derive them, same as single categories of social media user profiling approaches with multiple categories of explainability techniques.

[A05] Select classification procedure type. The qualitative procedure is applied on the results of the systematic literature reviews (SLR) (Kitchenham & Charters, 2007), which are to be conducted on the terms identification step in chapter 5. Subsequently, credit scoring components, social media user profiling approaches applicable to credit scoring, and the respective explainability techniques are categorized and put in relation to each other in chapter 6.

[A06] Identify the sources of information. The information sources used in the development of the proposed taxonomy are selected and described during the systematic literature reviews (SLR) (Kitchenham & Charters, 2007) for terms identification in chapter 5.

5 Taxonomy terms identification

Terms identification is conducted by focusing on the main three parts respectively resulting from the RQ1 as follows.

What are the techniques to provide explainability of social media user profiling approaches in credit scoring?

Hence, as first, in the subsection 5.1. the credit scoring model components are identified. Next, in the subsection 5.2. the social media user profiling approaches applicable to credit scoring are identified. Finally, in the subsection 5.3. the explainability techniques of the identified social media user profiling approaches in credit scoring are identified.

Following the taxonomy construction methodology, both respective activities [A07] and [A08] of the identification and extraction phase are conducted in each of the following subsections.

5.1 Credit scoring model components

The determination of the components affecting the credit score of consumer credit applicants is conducted by the systematic literature review (SLR) (Kitchenham & Charters, 2007). The focus is on the following research questions:

CH5-SLR1-RQ1. What are the traditional components affecting the credit score of consumer credit applicants?

CH5-SLR1-RQ2. What are the additional components potentially derivable from the social media data affecting the credit score of consumer credit applicants?

The search process, the inclusion and the exclusion criteria, the data collection, and the data analysis of the conducted SLR are explained as next. In the concluding part of this section the discussions to the defined in this section research questions are provided.

The search process is a manual search of suitable academic publications and literature using CatalogPlus, the comprehensive TU Wien academic research portal (TU CatalogPlus Search,

2020). In order to most comprehensively cover the components of credit scoring models the decision is made to use the broadest search string possible. The obvious resulting disadvantage is having to deal with a large amount of search results to process. On the other side, the high importance to identify the variety of possible credit scoring model components justifies such decision. The used search string is consequently as follows:

"credit scoring model"

Table 5. Credit scoring model components SLR search string

The inclusion criteria are as follows.

CH5-SLR1-IC1. Published in a peer-reviewed journal.

CH5-SLR1-IC2. Focus on credit scoring model(s) for consumer credits.

CH5-SLR1-IC3. Components that impact consumer credit scoring model are addressed.

The exclusion criteria are as follows.

CH5-SLR1-EC1. There is no access through TU Wien student account.

CH5-SLR1-EC2. The language of the publication is neither English nor German.

During the data collection stage, a search protocol was used, the summary of which is in Table 6.

Search query	Results	IC1	IC1, IC2	EC1	EC2	IC3	Selected
"credit scoring model"	1157	423	131	5	7	40	40

Table 6. Credit scoring model components SLR data collection

From each selected publication its bibliographic information (title, authors, publication year) is extracted. The obtained data (without duplicates) is tabulated to show the extracted information, as Table 14 in Appendix A. illustrates. The content of each publication is then analysed to provide the answers to the defined research questions CH5-SLR1-RQ1 and CH5-SLR1-RQ2.

[A07] Extract all terms. All extracted credit scoring model components are listed with their respective sources in Appendix B. Credit scoring model components extracted terms.

[A08] Perform terminology control. Inconsistencies in the extracted data, e.g., referral to the same credit scoring model components by different authors by different name, are accordingly removed, with respective details provided in Appendix C. Credit scoring model components terminology control.

The resulting list of the credit scoring model components is provided in Appendix D. Credit scoring model components.

5.2 Social media user profiling approaches

The determination of the approaches for social media user profiling applicable in credit scoring is conducted by the systematic literature review (SLR) (Kitchenham & Charters, 2007). The focus is on the following sole research question:

CH5-SLR2-RQ1. What are the approaches for social media user profiling applicable to credit scoring?

The search process, the inclusion and the exclusion criteria, the data collection, and the data analysis of the conducted SLR are explained as next. In the concluding part of this section the discussions to the defined in this section research question are provided.

The search is conducted as a manual search on CatalogPlus (the comprehensive TU Wien academic research portal (TU CatalogPlus Search, 2020)). In order to most completely cover the approaches for social media user profiling applicable to credit scoring, the decision is made to use the broadest search string possible. The obvious resulting disadvantage is having to deal with a large amount of search results to process. On the other side, the central importance of social media user profiling approaches applicable to credit scoring for this thesis justifies such decision. The used search string is consequently as follows:

“user profiling”

Table 7. Social media user profiling approaches SLR search string

The inclusion criteria are as follows.

CH5-SLR2-IC1. Published since the 1st of January 2010.

CH5-SLR2-IC2. Published in a peer-reviewed journal.

CH5-SLR2-IC3. Focus on social media user profiling.

Exclusion criteria:

CH5-SLR2-EC1. There is no access (through TU Wien student account).

CH5-SLR2-EC2. The language of the publication is neither English nor German.

CH5-SLR2-EC3. None of the identified credit scoring model components is addressed.

During the data collection stage, a search protocol was used, the summary of which is in Table 8.

Search query	Results	IC1	IC1, IC2	EC1	IC1, IC2, IC3	EC2	EC3	Selected
"user profiling"	2947	1765	725	0	85	0	0	85

Table 8. Social media user profiling SLR data collection

From each selected publication its bibliographic information (title, authors, publication year) is extracted. The obtained data (without duplicates) is tabulated to show the extracted information, as Table 15 in Appendix E. Social media user profiling SLR search results illustrates. The content of each publication is then analysed to provide the answers to the defined research question CH5-SLR2-RQ1.

[A07] Extract all terms. All extracted social media user profiling approaches applicable to credit scoring are listed with their respective sources in Appendix F. Social media user profiling approaches extracted terms.

[A08] Perform terminology control. Inconsistencies in the extracted data, e.g., referral to the same social media user profiling approach by different authors by different name, are accordingly removed, with respective details provided in Appendix G. Social media user profiling approaches terminology control.

The resulting list of the social media user profiling approaches applicable to credit scoring is provided in Appendix H. Social media user profiling approaches.

5.3 Explainability techniques for social media user profiling

The identification of the existing explainability techniques for social media user profiling approaches applicable in credit scoring is conducted by the systematic literature review (SLR) (Kitchenham & Charters, 2007). The focus is on the following sole research question:

> CH5-SLR3-RQ1. What are the techniques for explainability of social media user profiling approaches applicable to credit scoring?

The search process, the inclusion and the exclusion criteria, the data collection, and the data analysis of the conducted SLR are explained as next. In the concluding part of this section the discussions to the defined in this section research question are provided.

The search is conducted as a manual search on CatalogPlus (the comprehensive TU Wien academic research portal (TU CatalogPlus Search, 2020)). In order to most completely cover the explainability techniques for social media user profiling approaches applicable to credit scoring, the decision is made to use the broadest search string possible. The obvious resulting disadvantage is having to deal with a large amount of search results to process. On the other side, the central importance of explainability techniques of social media user profiling approaches applicable to credit scoring for this thesis justifies such decision. The used search string is consequently as follows:

"explainability"

Table 9. Explainability of social media user profiling approaches SLR search string

The inclusion criteria are as follows.

CH5-SLR3-IC1. Published in a peer-reviewed journal.

CH5-SLR3-IC2. Focus of the publication is on explainability technique(s).

CH5-SLR3-IC3. At least one of the determined social media user profiling algorithms applicable to credit scoring is categorized (i.e., assigned to a concrete explainability technique or general class of explainability techniques).

Exclusion criteria:

CH5-SLR3-EC1. There is no access (through TU Wien student account).

CH5-SLR2-EC2. The language of the publication is neither English nor German.

During the data collection stage, a search protocol was used, the summary of which is depicted in Table 10.

Search query	Results	IC1	IC1, IC2	EC1	EC2	IC1, IC2, IC3	Selected
"explainability"	1318	474	61	0	0	33	33

Table 10. Explainability of social media user profiling approaches SLR data collection

From each selected publication its bibliographic information (title, authors, publication year) is extracted. The obtained data (without duplicates) is tabulated to show the extracted information, as Table 16 in Appendix I. Explainability techniques SLR search results illustrates. The content of each publication is then analysed to provide the answers to the defined research question CH5-SLR3-RQ1.

[A07] Extract all terms. All extracted explainability technique for social media user profiling approaches applicable to credit scoring are listed with their respective sources in Appendix J. Explainability techniques extracted terms.

[A08] Perform terminology control. Inconsistencies in the extracted data, e.g., referral to the same explainability technique for social media user profiling approach by different authors by different name, are accordingly removed, with respective details provided in Appendix K. Explainability techniques terminology control.

The resulting list of the explainability techniques for social media user profiling approaches applicable to credit scoring is provided in Appendix L. Explainability techniques.

6 Taxonomy construction

Taxonomy construction proceeds by specifying the main dimension, identifying its categories along all of the defined taxonomy levels of credit scoring model components, social media user profiling approaches for credit scoring, and explainability techniques, the relationships between categories, and hence resulting in a classification scheme for the defined subject matter. These activities are respectively described in the subsections 6.1., 6.2., and 6.3. In the conclusion of this chapter, in the subsection 6.4., the guidelines for using and update the taxonomy are defined.

6.1 Taxonomy dimensions

[A09] Identify and describe taxonomy dimensions. For the selected taxonomy structure type, one single dimension is identified at the top (i.e., root of the taxonomy) along which the defined subject matter is classified:

Explainability technique for social media user profiling approaches in credit scoring

6.2 Taxonomy categories

[A10] Identify and describe categories of each dimension. Categories are identified in the following subsections aligned with the split to the three levels as discussed in chapter 5, namely credit scoring model components categories, approaches of social media user profiling categories, and explainability techniques of approaches for social media user profiling categories.

6.2.1 Categories of credit scoring model components

The extracted terms of credit scoring model components, as identified in the subsection 5.1., are categorized by following a hybrid approach (Usman, 2015, p. 124) that combines traditional top-down and bottom-up approaches (Broughton, 2015) in the following way: the initial set of credit scoring model components categories is extracted and terminology controlled from the selected publications of the SLR on credit scoring model components in the subsection 5.1. (i.e., top-down); identified single credit scoring model components are successively assigned to matching

categories, creating new category in case no matching category for a particular term exists (i.e., bottom-up).

As the result, the identified categories for the credit scoring model components are as follows:

- Bank-borrower relationship
- Collateral characteristics
- Credit applied for
- Credit card(s) data
- Credit history
- Demographic data
- Employment status
- Financial indicators
- Look-a-likes
- Psychological variables
- Semiometric space
- Social network data
- User-generated content

These categories are described in the following subsections in details, providing the information on which of the extracted terms belong to which of the identified categories.

Bank-borrower relationship

Bank-borrower relationship includes the following components: *balance of the current account, balance on checking account(s), balance on savings (deposit) account(s), bank accounts, banking activity, length of relationship (years at bank), life insurance policies, number of existing credits at the bank, proximity to bank, type of account(s), charge card, cheque card, overdraft, relation with bank, relation with other banks*.

Bank-borrower relationship is distinguished as a separate category of credit scoring model components among others by (Gan, Li, Wang, & Kao, 2012, p. 337).

(Gan, Li, Wang, & Kao, p. 4) provided an extensive elaboration on the potential impact of bank-borrower relationship for credit scoring. According to their findings, there is an evidence that a good relationship between borrower and bank could result in more favourable credit conditions, such as lower credit rate, lower required collateral, greater availability of funds, etc. They

underscore a good relationship in first place by the length of the relationship, proximity to bank, and an overall usage of financial services. Provided justification of a good bank-borrower relationship importance is based on facilitated monitoring and screening capabilities to be able to better assess the applicants' financial status, hence to overcome problems of asymmetric information, and as the result potentially lowering the probability of default.

Similar to relationship with bank applied at, also relationship with other banks is of a certain importance for credit scoring (Abdou H. A., 2009, p. 5). Although the information about present accounts and the overall banking activity at other banks is naturally more difficult to obtain, such data might be beneficial especially in case of having no account at bank where applied for credit (Janeska, Taleska, & Sotiroski, 2014).

Collateral characteristics

Collateral characteristics include the following components: *collateral, value of property, floor space of the property, current mortgage, current car credit, car asset, real estate, most valuable available asset*.

Collateral characteristics is distinguished as a separate category of credit scoring model components among others by (Chi & Hsu, 2012) (Kiss, 2003, p. 96) (Siami, Gholamian, & Basiri, 2014).

Collateral, while helping credit applicant to secure the credit, provides for lender the possibility to mitigate the risk of borrower's default on provided credit by repossessing the collateral, hence collateral area identified to be similarly important as e.g., borrower's age, borrower's education, or borrower's occupation (Chi & Hsu, 2012, p. 6). As the result, also other collateral characteristics often play an important role in credit scoring models, such as car asset (Shen, Wang, & Shen, 2020, p. 416), or any other most valuable available asset (Zeng, 2017, p. 7748).

Credit applied for

The attributes of interest of credit applied for include: *amount of the monthly instalment, borrowing interest rate, credit conditions, credit purpose, credit repayment type, credit type, credit amount, credit-to-value ratio, credit duration, own contribution (as per credit purpose), monthly repayment burden*.

Characteristics and conditions of credit applied for is distinguished as a separate category of credit scoring model components among others by (Baklouti, 2014a, p. 199) (Siami, Gholamian, & Basiri, 2014).

The attributes of the credit applied for are usually very important for credit scoring models (Abid, Masmoudi, & Zouari-Ghorbel, 2016) (Chi & Hsu, 2012) (Chuang & Lin, 2009) (Dimitriu, Avramescu, & Caracota, 2010) (Kim & Sohn, 2004). For example, in case of mortgage, the odds for a credit to default are much lower if the purpose of credit is for self-living or repair than if the purpose is investment (Chi & Hsu, 2012, p. 6), and own contribution of at least 25% of the investment value is desired, with the best score assigned for own contribution of 50% and more (Dimitriu, Avramescu, & Caracota, 2010, p. 6). The lower amount of the requested credit, especially in conjunction with a shorter credit duration, the lesser financial loss would default on such credit cause, hence increasing the probability of the credit being approved (Kim & Sohn, 2004). The credit amount is identified to be similarly important as e.g., the duration of the present employment (Chuang & Lin, 2009, p. 7). The monthly repayment burden determined to have a direct effect on the potential default payment, hence high predictive power for the detection of customers' credit default (Abid, Masmoudi, & Zouari-Ghorbel, 2016).

Credit card(s) data

Credit card(s) data of interest include: *credit card status, type of credit card(s), preferred credibility limit (credit cards), average consumption and maximum consumption of credit card in the past 6 months, time from the first credit card to the current time (i.e., credit time limit class), credit card repayment speed, credit card history overdue total days, credit card overdue amount*. Information on credit cards is distinguished as a separate category of credit scoring model components among others by (Zhang, Zeng, Chen, & Zhang, 2020, p. 4).

Credit card(s) attributes are distinguished in a specific set of components (potentially) influencing credit scoring models, not least as the result of their nowadays' widespread use (Zhang, Zeng, Chen, & Zhang, 2020, p. 4). Among various credit card(s) attributes, type of credit card is of particular interest for credit scoring (Anderson, 2019, p. 352).

Credit history

The elements of interest from credit history include: *forced early repayment status, active early repayment status, defaulter/non-defaulter (past defaults), number of previous credits, previous*

credits defaults, other installment plans (credits), average credit repayment time, total historical credit amount, overdue repayment, number of times and overdue duration, number of current credits, number of institutions with present credits from, amount and time of deferred repayment of specific business, amount and time of repayment of specific business, delinquency status in the last 3-6-12 months, time of inquiry by the organization and corresponding reasons, time from the first credit to the current time (i.e., credit time limit class), credit history.

Credit history (namely as past payments characteristics, i.e., repayment history) is distinguished as a separate category of credit scoring model components among others by (Chi & Hsu, 2012) (Zhang, Zeng, Chen, & Zhang, 2020, p. 4).

Credit history is identified as a very important component that usually strongly influences credit scoring models (Chuang & Lin, 2009; Kim & Sohn, 2004; Zhang, Zeng, Chen, & Zhang, 2020; Janeska, Taleska, & Sotiroski, 2014; Shi, Zhang, & Qiu, 2013; Hsieh, 2005) (Zhou, Lai, & Yu, 2009; Akkoç, 2012). Based on data about previous credits, for example, the lowest points are assigned in case of problems with credits in the past, and the second lowest points if the applicant had no credits in the past, as opposed to the highest points given in case of all of the past credits paid and on time, and the second highest points if the applicant is paying current credits on time (Tomczak & Zięba, 2015, p. 1791). On the other side, some components are although found to be used in credit scoring models, their usefulness is doubted, such as the delinquency status in the last 3-6-12 months (Chi & Hsu, 2012, p. 2655).

Demographic data

Demographic attributes include: *gender, education, marital status, housing, present residence, residence region, current electoral roll category, Age, number of children (under 16), time at current/previous residence, number of dependents, weeks since last county court judgement, television area code, years on electoral roll at current address, Number of guarantors, other debtors, co-applicant information, spouse's income, credit status of guarantor, ethnicity, home duration, identity certification, number in household, socio-demographic data, telephone*.

Borrower demographic data is distinguished as a separate category of credit scoring model components among others by (Chi & Hsu, 2012) (Abbod & Radi, 2018, p. 616) (Baklouti, 2014a, p. 199) (De Cnudde, et al., 2019) (Guo, et al., 2016).

Credit scoring models are often seen to be incomplete without incorporating at least some of the demographic data of credit applicants (Chi & Hsu, 2012) (Abbod & Radi, 2018, p. 616) (Baklouti, 2014a, p. 199) (De Cnudde, et al., 2019) (Guo, et al., 2016). At the same time, single components of demographic data are identified as lesser or stronger predictors for borrowers' credit default. On the one hand, credit applicant's age is often used in credit scoring models (Abbod & Radi, 2018) (Abdou H. A., 2009) (Abdou, Alam, & Mulkeen, 2014) (Abid, Zaghdene, & Masmoudi, 2017) (Akkoç, 2012) (Anderson, 2019), and found to be particularly useful (Chi & Hsu, 2012, p. 2655). Although the exact boundaries of different age categories are usually set differently in different credit scoring models, certain pattern of giving points depending on age is observable, for example: highest points for applicants of age between 35 and 60 (Tomczak & Zięba, 2015, p. 1791) vs. of age between 30 and 50 (Janeska, Taleska, & Sotiroski, 2014, p. 53), second highest points for applicants of age between 18 and 35 (Tomczak & Zięba, 2015, p. 1791) vs. of age between 20 and 30 (Janeska, Taleska, & Sotiroski, 2014, p. 53), and the lowest points for applicants of age over 60 (Tomczak & Zięba, 2015, p. 1791) vs. of age over 50 (Janeska, Taleska, & Sotiroski, 2014, p. 53). Another important demographic attribute, namely level of education, is also identified to be a good default predictor often used in credit scoring models (Chi & Hsu, 2012, p. 2655). The arguments pro better educated borrowers are their more stable employment and higher income, thus lower default rate (Gan, Li, Wang, & Kao, 2012, p. 338). On the other hand, e.g., the impact of applicant's gender on credit default probability is rather disputed, ranging from dropping gender as evidently not useful (Chi & Hsu, 2012, p. 2655) to ample evidence suggesting that women default less frequently because of being more risk adverse (Gan, Li, Wang, & Kao, 2012, p. 338), not even considering the potential for discrimination if taking gender into account in credit scoring models, hence unlawfully to be used. Marital status, although affecting the credit applicant's responsibility, reliability, and maturity, is rather discovered to result in higher probability of default for married than single, since typically being related to the number of dependents hence financial pressure on the borrower (Gan, Li, Wang, & Kao, 2012, p. 338). Finally, possible impact of residence on credit scoring is justified by the consideration of people of similar wealth tending to live in the same neighbourhood, i.e., residence potentially indicating credit applicant's level of financial wealth and status (Gan, Li, Wang, & Kao, 2012, p. 338).

Employment status

The elements of interest regarding employment status include: *company, occupation, job status, business sector (industry), company type, company size, occupation group, mode of work, mode of income, job title/position, current income, work seniority, job experience, time at previous job, total working duration, job*.

Employment status is distinguished as a separate category of credit scoring model components among others by (Abbod & Radi, 2018, p. 616).

Occupation is widely believed to be highly correlated with the borrower's income, hence commonly used in credit scoring models (Gan, Li, Wang, & Kao, 2012, p. 338). The type of employment could indicate income and financial stability, with e.g., fixed salary employees having lower default risk, and unemployed credit applicants considered as not creditworthy (Abid, Masmoudi, & Zouari-Ghorbel, 2016, p. 958). Employment duration and concrete jobs are another important components of credit scoring models, e.g., borrowers with longer and more reliable employment history, similar to borrowers with professional jobs, are less likely to default (Chuang & Lin, 2009, p. 1691) (Gan, Li, Wang, & Kao, 2012, p. 338).

Financial indicators

The components regarding financial indicators of credit applicant include: *debt-to-income ratio, monthly expenses (outgoings, spending monthly), life insurance policies, number of searches in last 6 months, capacity, capital, character, credit certificate, debt ratio, expenses, financial credibility, outstanding credit*.

Financial indicators are distinguished as a separate category of credit scoring model components among others by (Abbod & Radi, 2018, p. 616).

Monthly income versus monthly expenses, debt-to-income ratio, and other financial indicators are naturally considered as some of the key determinants in credit scoring (Chi & Hsu, 2012) (Abdou, Alam, & Mulkeen, 2014). E.g., outstanding credit impacts the borrower's likelihood of default since the greater the amount of outstanding credits the higher the chances to default (Abid, Masmoudi, & Zouari-Ghorbel, 2018, p. 958).

Look-a-likes

Look-a-likes of credit applicant: *interest-based look-a-likes, relational look-a-likes*.

Look-a-likes is distinguished as a separate category of credit scoring model components among others by (De Cnudde, et al., 2019).

(De Cnudde, et al., 2019) study the potential of complementing traditional credit scoring data with Facebook data in the microfinance setting, with one of the main focuses on the relationship between users in form of look-a-likes (LALs). Their determination is conducted based on socio-demographic data (such as age, place of residence, education), interest data (such as liked pages or companies worked for), and social network data (friendship connections). Additionally, LALs are split into relational and interest-based LALs, defined as follows (De Cnudde, et al., 2019).

- Interest-based LALs refer to people that explicitly manifest similar interests (e.g., liking a Facebook page or joining a specific group). Used Facebook data indicating interest-based LAL relationship: persons liking a page on Facebook, persons liking a category of a page on Facebook, persons joined in a group on Facebook, persons going to specific educational institutions, persons working for employers, persons holding employment positions or business titles.
- Relational LALs refer to people that are similar to one another by inspecting the interactions between users (i.e., in terms of text, links, photos, videos being shared on someone's wall, tagged, commented on, liked). Used Facebook data indicating relational LAL relationship: persons commenting on a status, persons mentioned in a picture, persons mentioned in a link, persons mentioned in a status, persons mentioned in a video, persons liking an item (video/status/photo/comment), persons giving/receiving comments to/from each other, persons mentioning one another in one of their photos, persons mentioning one another in one of their links, persons mentioning one another in one of their statuses, persons mentioning one another in one of their videos, persons liking each other's video/status/photo/comment, persons giving/receiving comments to/from each other, persons mentioning one another in one of their photos, persons mentioning one another in one of their links, persons mentioning one another in one of their statuses, persons mentioning one another in one of their videos, persons liking each other's video/status/photo/comment.

(De Cnudde, et al., 2019) conclude interest-based data yielding better results than the person's social network data. Overall, the developed model built solely on interest data is even not significantly worse than the model that uses all of the available data, hence underscoring the high potential of utilizing Facebook data in credit scoring (De Cnudde, et al., 2019).

Psychological variables

Borrower's psychological traits re: *miscalibration, better-than-average, illusion of control, emotional Intelligence*.

Psychological variables are distinguished as a separate category of credit scoring model components among others by (Baklouti, 2014a).

(Baklouti, 2014a) study the potential of using borrowers' psychological traits for predicting future credit defaults. The focus is on overconfidence and emotional intelligence, which were measured based on the respective questionnaire. Overconfidence, specified as certain behaviour reflecting one's tendency to overestimate one's ability and chances for success, the probability of gaining positive outcomes, and the accuracy of possessed knowledge, is further split into miscalibration, better-than-average, and illusion of control. Hence, assessed borrower's psychological traits are the following (Baklouti, 2014a, p. 200).

- Miscalibration, i.e., overestimation of the person's capacities to make the right predictions, measured to capture the subject's subjective confidence of their knowledge.
- Better-than-average, i.e., bias of individuals feeling better than others.
- Illusion of control, i.e., measurement of the respondents' certainty in their ability to master and to predict difficult-to-control, future events.
- Emotional intelligence, i.e., measured level of emotional intelligence.

(Baklouti, 2014a) confirm the existence of the prevalence of psychological traits and their specific effect on decision making, namely behavioural traits characterized by overconfident behaviour, widespread use of heuristics, and emotional intelligence. In particular, psychological traits highlighted to relate to investment behaviour, psychological biases such as overconfidence to lead to incorrect information processing and other major misperceptions, emotional intelligence to play an important role in helping people identify and interpret various cues in life, increasing chances for more rational decision-making, etc. Hence, (Baklouti, 2014a) conclude that borrowers' psychological traits constitute a major information source in predicting creditworthiness.

Semiometric space

Borrower projection onto the semiometric space: *duty/pleasure, attachment/detachment, sublimation/materialism, idealization/pragmatism, humility/sovereignty*.

Semiometric space is distinguished as a separate category of credit scoring model components among others by (Liberati & Camillo, 2018).

(Liberati & Camillo, 2018) analysed the relationship between personality traits and financial behaviour, focusing in particular on the potential to lower credit risk in scoring models by evaluating borrower's projection on the so-called semiometric space.

The idea behind the Sémiométrie is the affirmation that terms and words are associated with emotional meanings, i.e., Sémiométrie is the collection of words marked by survey respondents in terms of sensation (pleasant or unpleasant), with selection of the words conducted based on the criteria of non-consensuality, semantic uniqueness, semantic stability, and evocative power (Liberati & Camillo, 2018, p. 1995). Then, the Principal Component Analysis (PCA) is applied to scale and synthesize the dimensions of the multidimensional table of responses, resulting in a total of 210 principal components, 6 of which interpreted and described as follows (Liberati & Camillo, 2018, p. 1996).

- Participation reflects the attitude of the respondents to the questionnaire.
- Duty/pleasure contrasts two different ideas, namely focus on following the rules versus on enjoyment.
- Attachment/detachment discloses differences between those that tend to create links with objects or other people and those that are self-sufficient.
- Sublimation/materialism describes the usual contrariness between body and soul.
- Idealization/pragmatism refers to the opposition between the needs to overreaching and to dream and the needs to understand and to act in a rational world.
- Humility/sovereignty describes the opposition between ordinary people and those standing out from the crowd.

Excluding the first principal component of participation, which is only related to the applied methodology, hence the total of five semiometric factors are subsequently used by (Liberati & Camillo, 2018) to construct a subspace on which borrowers are projected. At the same time, by conducting projections it is not expected to have correlation with the rational meaning of single selected factors, i.e., projections are meant as expressions of emotional and subjective meanings linked to the personal experiences, cultural environment and emotional feelings of the respondent. In other words, constructed semiometric space is solely an average space of connotations of the collective unconscious.

(Liberati & Camillo, 2018) confirm credit scoring model accuracy improvement when default risk is estimated under the consideration of non-economic variables such as personal values and personality traits evaluated based on the borrower's projection onto the semiometric space.

Social network data

Social network data of credit applicant: *social network, user social network*.

Social network data is distinguished as a separate category of credit scoring model components among others by (Guo, et al., 2016) (De Cnudde, et al., 2019).

(Guo, et al., 2016) considered among others the user social network, namely relationships of friends, followers, followees, and ego-network structures, in their research on the possibilities to leverage social media data for personal credit scoring. The study by (De Cnudde, et al., 2019) on the potential of complementing traditional credit scoring data with Facebook data in the microfinance setting, besides look-a-likes, focuses on the relationships between users in form of friends, and so-called best friends forever (BFFs), defined as follows (De Cnudde, et al., 2019).

- Friends are those explicitly marked as friends on Facebook. Used Facebook data indicating friendship relationship: befriending one another.
- BFFs are those explicitly marked as friends on Facebook, who are also (frequently) interacting with one another (e.g., being tagged on together on picture(s), commenting on each other posts, status updates, etc.). Used Facebook data indicating BFF relationship: friends giving/receiving comments to/from one another, friends mentioning one another in one of their photos, friends mentioning one another in one of their links, friends mentioning one another in one of their statuses, friends mentioning one another in one of their videos, friends liking each other's video/status/photo/comment, friends having any kind of interaction.

(De Cnudde, et al., 2019) conclude BFFs having higher credit default predicting power than the person's friends, overall underlining the high potential of utilizing Facebook data in credit scoring (De Cnudde, et al., 2019).

User-generated content

Borrower's social media data regarding user-generated content: *user-generated content*.

User-generated content is distinguished as a separate category of credit scoring model components among others by (Guo, et al., 2016).

(Guo, et al., 2016) dedicated their research to the possibilities of leveraging user-generated social media data for personal credit scoring, i.e., learning social media users' creditworthiness labels in

a comprehensive and efficient way. Their study is conducted in the micro-blogging setting with its respective specific features available. Based on proposed social-data-based credit scoring principles and gained credit-related insights from empirical observations of test data, (Guo, et al., 2016) elaborated prediction features extracted from the social media users' demographics data, tweets, and networks.

Proposed social-data-based credit scoring principles are the following.

- Capacity-principle. "Good credit users are more willing to share moments about their personal lives on the social platforms. Some of these moments suggest that they are capable of paying back the credit debt in time. Their economic capacity is usually very stable for meeting their payments" (Guo, et al., 2016, pp. 7-8).
- Character-principle. "Good credit users are more likely to exhibit characteristics indicating that they are content contributors rather than consumers on social media. They also have the characteristics of being prudent and responsible, reflected from their writing styles and content qualities" (Guo, et al., 2016, p. 8).
- Conditions-principle. "Good credit users maintain good mental and physical conditions, ensuring that no external misfortunes like unemployment or ill-health happen to them in the future. Good health improves one's ability to repay the credit to at least some degree" (Guo, et al., 2016, p. 8).

Gained credit-related insights from empirical observations of social media users' test data are the following (Guo, et al., 2016, pp. 8-10).

- Economic Stability. Good credit users are expected to possess a stable income, and to have a stable future work prospect. Additionally, users with more tweets about work are found to much more unlikely default on credit. Examples of features extracted from social media data that correspond to economic stability insight: age, occupation types.
- Experienced Employee. Good credit users tend to work at certain jobs for a relatively long period of time, being experts in certain areas, holding higher (i.e., more senior) positions. Employment at famous companies further increases the probability for a more stable income. Examples of features extracted from social media data that correspond to experienced employee insight: number of years since the user starts his or her career, number of companies where the user has worked.
- Well-Educated. A good education or a high academic degree lowers credit default risk. One's education can be also indirectly concluded from language style and tweet topics. Examples of features extracted from social media data that correspond to well-educated insight: education level, sentiment vocabulary (e.g., vulgar language).

- Creative Poster. Good credit users are determined to spend more time posting and sharing their personal affairs rather than retweeting or tweeting about news, reviews, quotes, old sayings, etc. Moreover, good credit users have more positive attitude toward life and work. Examples of features extracted from social media data that correspond to creative poster insight: usage of emoticons, average length of retweet chains.
- Healthy Lifestyle. Bad credit users are recognized to tweet more during early hours in contrast to good credit users that tend to tweet more during daylight or evening hours, concluding the posting time distribution to be a strong indicator of users' activity intensity during days and nights. It is also observed that bad credit users seem to talk more often about suffering from different illnesses such as flu or insomnia. Examples of features extracted from social media data that correspond to healthy lifestyle insight: fraction of tweets published at each hour during the day, sentiment polarity distribution.
- Prudence and Responsibility. Good credit users determined to tend to be prudent and responsible, being more concerned about the rules of modern society, hence more likely keep promises and thus maintain creditworthiness. Examples of features extracted from social media data that correspond to prudence and responsibility insight: number of duplicate tweets, aggregated features of one-hop neighbours' degree features.

Extracted prediction features (with custom high-level features) are the following (Guo, et al., 2016, pp. 14-23).

- Demographic features: length of the screen name, number and proportion of alphabetic characters in the screen name, number and proportion of numerical characters in the screen name, number and proportion of symbol characters in the screen name, gender and Age of the user, whether the user's identity is verified by Weibo or not, education level of the user, provinces where the user lives, number of companies where the user has worked, number of years since the user starts his or her career, whether the company the user works in is renowned or not, number of years and months since the user joined Weibo, active level of the user.
- Tweet features: number and fraction of retweets of a user's tweets, number and fraction of retweets with no comments, average depth of retweet chains, maximum depth of retweet chains, depth deviation of retweet chains, number of emoticons/mentions in users' tweets, standard deviation of number of emoticons/mentions in a user's tweets, average number of emoticons per tweet, fraction of tweets that contain emoticons/mentions, fraction of tweets at each of 24 hours of a day, number and fraction of tweets whose sentiment polarities are, positive/negative/neutral respectively, deviation of the sentiment polarity values among users' tweets, number of positive/negative

sentiment word occurrences in users' tweets, fraction of positive/negative sentiment words in users' tweets.

- Network features: number of followers, number of friends, fraction of followers that are also followees, fraction of followees that are also followers, fraction between number of followers and followees, aggregated values of a user's one-hop neighbours' network features, betweenness centrality, PageRank values.
- High-Level features: features derived from ngram features using Logistic Regression, features derived from ngram features using Naive Bayes, features derived from topic distributions using Logistic Regression, features derived from topic distributions using Naive Bayes, features derived from topic distributions using Decision Tree, features derived from demographic features with different classifiers, features derived from tweet features with different classifiers, features derived from network features with different classifiers.

6.2.2 Categories of social media user profiling approaches

The extracted terms of approaches for social media user profiling in credit scoring, as identified in the subsection 5.2., are categorized by following a hybrid approach (Usman, 2015, p. 124) that combines traditional top-down and bottom-up approaches (Broughton, 2015) in the following way: the initial categories of approaches for social media user profiling in credit scoring are extracted and terminology controlled from the selected publications of the SLR on approaches for social media user profiling in credit scoring in the subsection 5.2. (i.e., top-down); identified single approaches for social media user profiling in credit scoring are successively assigned to matching categories, creating new category in case no matching category for a particular term exists (i.e., bottom-up).

As the result, the identified categories for the approaches for social media user profiling in credit scoring are as follows:

- Artificial Neural Networks
- Clustering
- Decision trees
- Dimensionality reduction
- Ensemble learning
- Graph theory algorithms
- Linear models
- Natural Language Processing

- Nearest neighbour models
- Probabilistic and statistical models
- Social recommender systems
- Social semantic web
- Support vector machines

Mostly all of the social media user profiling approaches extracted from the selected publications belong or are closely related to machine learning (ML). Strongly simplified, the core idea of ML is based on the assumption that outcome y relates to input x in form of a simple equation $y = f(x)$, and the task of ML is to computationally find such function h that most closely approximates the true function f (Russell & Norvig, 2010, p. 695). Usually, the function h is called a hypothesis, the input x is generally in form of a vector, and the output y is a single value, whereby values in x and of y don't necessarily need to be numbers, but can be any values (Russell & Norvig, 2010, p. 695). Furthermore, the differentiation between supervised, unsupervised, and semi-supervised learning is common as by the following (Russell & Norvig, 2010, pp. 694-695).

- Supervised learning: existing input-output pairs serve as the basis to learn function that maps input to output.
- Unsupervised learning: patterns in input data are learnt without the corresponding output being provided.
- Semi-supervised learning: both labelled (i.e., input-output pairs) and unlabelled (i.e., no corresponding output to given input) data has to be dealt with to determine the best hypothesis.

The identified categories are described in the following subsections in details, providing the information on which of the extracted terms belong to which of the categories.

Artificial Neural Networks

Approaches: *attention mechanism, autoencoders and perceptrons, bi-directional gated recurrent unit (biGRU, type of RNN), Bidirectional LSTM (BiLSTM), CNN, CNN ResNet-50, CNN to classify images (map to KG), CNN-RNN, combined perceptron with Bayes model, compositional recurrent neural network, deep learning, deep neural network, deep neural networks (bi-GRU layer, hierarchical attention layer, BiRNN, concatenation layer), DeepWalk, extend deep autoencoder with top-k semantic social information, Feature Refinement Layer, gated recurrent neural network (CNN-GRU), GRU structure, hierarchical attention network, Hierarchical Attention Transfer Network (HATN), hierarchical convolution neural network (CNN),*

ImageNet/GoogleNet, LSTM, LSTM (sentiment classifier), MLP (multi-layer perceptron), multi-granularity CNN, multi-modal deep belief network (DBN), multi-modal deep Boltzmann Machines (DBM), network representation learning (NRL), neural network model (social convolution attention neural network), neural networks, PGBN (Poisson Gamma Belief Network, a deep learning topic model), replicated softmax model, representation learning (feature learning), RNN, RNN based collaborative filtering, socially embedded visual representation learning (SEVIR), Softmax layer, text attention neural network model (TA-NN), multi-model user attribute model (mmUAM).

Artificial Neural Networks (ANN) are distinguished as a separate category of approaches for user profiling (Russell & Norvig, 2010, p. 727) to cover the other notions such as neural networks (Li, Yang, Xu, Wang, & Lin, 2019), deep learning (C C & Mohan, 2019) (Chen, Wang, Ren, Liu, & Lin, 2018), and single components and types thereof (Buraya, Farseev, & Filchenkov, 2018) (Li, Yang, Xu, Wang, & Lin, 2019).

The idea behind ANN is taken from neuroscience's concept of network of brain cells called neurons, which are schematically depicted on Figure 3 with their simplified mathematical model on Figure 4 (Russell & Norvig, 2010).

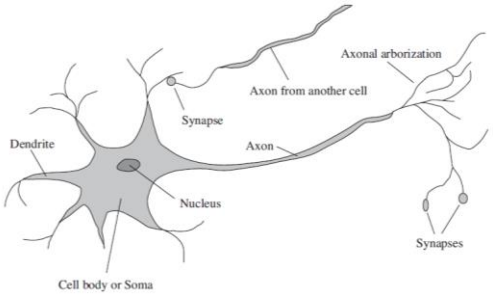

Figure 3. Schematic diagram of a neuron (Russell & Norvig, 2010)

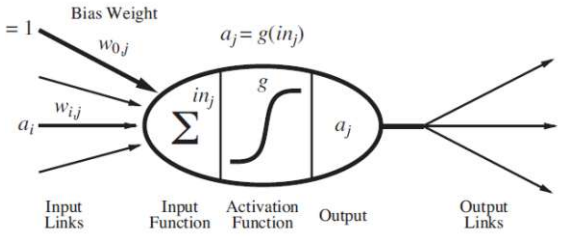

Figure 4. Mathematical model of a neuron (Russell & Norvig, 2010)

ANN is then a collection of neurons connected to each other in a specific manner, e.g., as depicted on Figure 5.

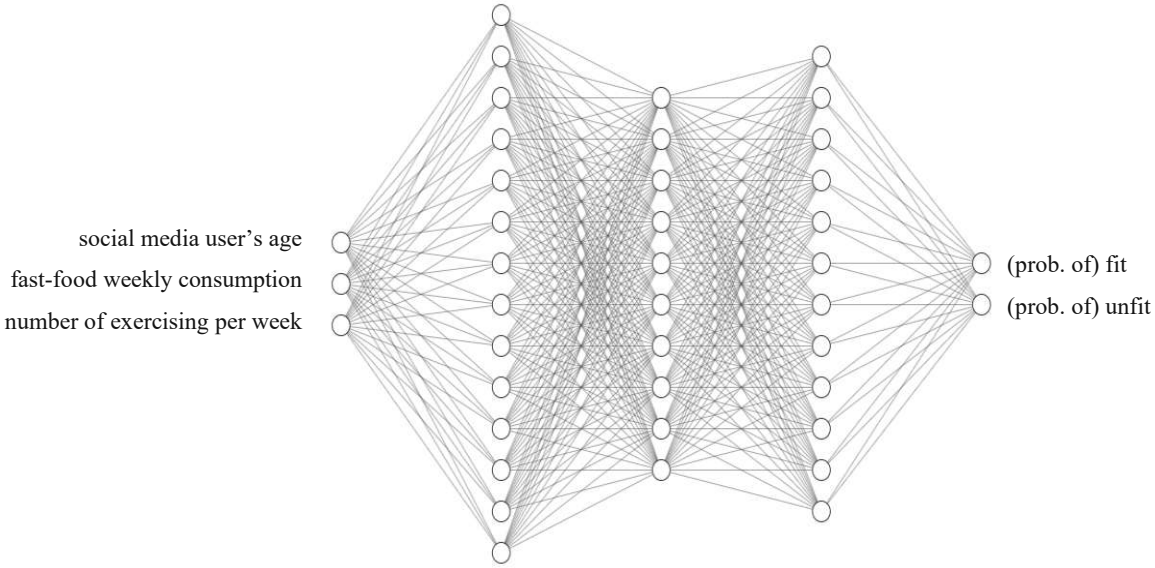

Figure 5. Fully-connected ANN with inputs, hidden layers, and outputs (Lenail, 2020)

As the result of a large variety of possibilities to construct neural networks, ANNs currently outperform most of the other approach types on the most different use cases, although at cost of their complexity (Russell & Norvig, 2010). In social media user profiling, ANNs are proved to be successfully appliable to a wide range of tasks such as image classification (Lully, Laublet, Stankovic, & Radulovic, 2018), extracting semantic information from texts and avatars (Li, et al., 2019), profiling users' preferences (C C & Mohan, 2019) and users' interests (Kang, Choi, & Lee, 2019), personality profiling (Buraya, Farseev, & Filchenkov, 2018), etc.

Clustering

Approaches: *affinity propagation clustering, centroid-based classification (text classification), clustering, Cosine Similarity, DBSCAN, dynamic user clustering topic model (UCT), fuzzy C-means algorithm for clustering, hierarchical clustering, k-means, Lucene Clustering, spectral clustering, applied standardization (clustering), constrained label propagation, label propagation, co-profiling algorithm, FCM, OKM, formal concept analysis (FCA), Levenshtein Similarity, Smith-Waterman Similarity*.

Clustering is distinguished as a separate category of approaches for user profiling among others by (Eke, Norman, Shuib, & Nweke, 2019).

Through clustering the aim is to group data into so-called clusters, without prior knowledge on the existence and labels of specific clusters (Russell & Norvig, 2010, p. 817). Hence, clustering

is an unsupervised approach, moreover the most common unsupervised learning task (Russell & Norvig, 2010, p. 694). *k*-means, one of the most popular clustering algorithms, requires initial selection of the number of clusters to construct, and then proceeds iteratively by assigning each data point to cluster with the closest so-called cluster centroid to that data point, recalculating cluster centroids after each step, and repeating this procedure until some stopping criterion is met (Cady, 2017, p. 145).

Consider an exemplarily task to cluster social media users based on number of days consuming fast food (per week) and number of days exercising (per week). Figure 6 contains the initial set of data points, which can subsequently be clustered into 3 clusters e.g., as illustrated on Figure 7.

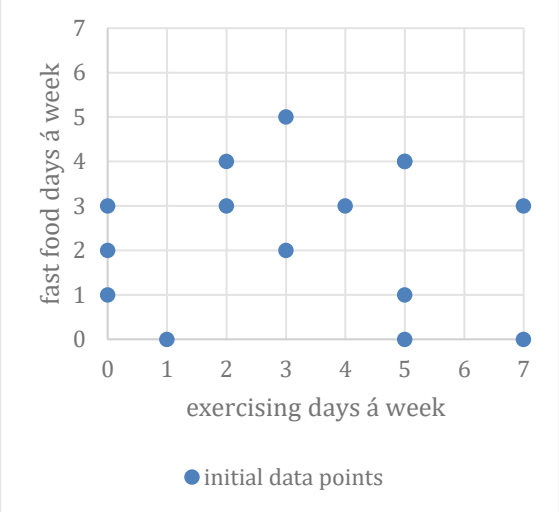

Figure 6. Initial data before clustering

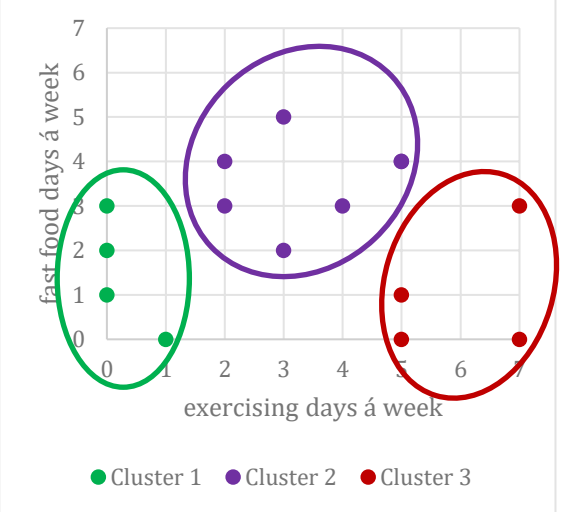

Figure 7. Data clustered into three clusters

In social media user profiling, clustering can be applied e.g., to divide tweets dataset into specific clusters such as daily chatter, conversations, sharing of information or links, and news reporting (Al-Qurishi, et al., 2018, p. 11181), to cluster users based on their behaviour into categories such as stars, chatters, socializers, concluders, observers (Barysheva, Petrov, & Yavorskiy, 2015, pp. 260-261), to cluster users' friends into different categories such as influential, less influential, and non-influential friends (Alshammari, Kapetanakis, Polatidis, Evans, & Alshammari, 2019, p. 97), and many others.

Decision trees

Approaches: *CART tree based model, condensed filter tree (CFT), decision tree, decision trees (J48, ADTree, REPTree), gradient boosted decision trees, rule-based, rule-based systems, tree-structured CRF (conditional random field)*.

Decision trees are distinguished as a separate category of approaches for user profiling among others by (Zhang & Bors, 2019, p. 216) (Guo, et al., 2016, p. 23) (Russell & Norvig, 2010, p. 697).

Decision trees are rather simple yet very powerful and widespread machine learning approach (Chen, Zhu, Guo, & Liu, 2014; Peng, Detchon, Choo, & Ashman, 2017; Guo, et al., 2016; Pang, Jiang, & Chen, 2013; Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016; Barbon, Igawa, & Bogaz Zarpelão, 2017; Faralli, Stilo, & Velardi, 2015; Zhang & Bors, 2019).

A decision tree represents a function that produces output to supplied input by performing a sequence of tests respectively on single values of input, proceeding in tree-structured manner, i.e., internal nodes contain tests and leaf nodes contain function output (Russell & Norvig, 2010, p. 698). Consider as an example a decision tree to decide whether a particular person is fit or unfit based on such available attributes as age, food consumption, exercising. This exemplarily decision tree is depicted on Figure 8.

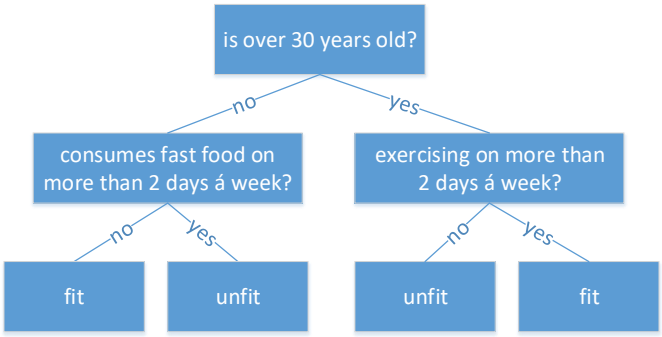

Figure 8. Decision tree to decide whether a person is fit or unfit

Decision trees are constructed from a set of example input-output pairs (i.e., supervised learning), aiming on the one side to find a good approximate solution, and on the other side to keep the tree as small as possible, what is usually reached by specific heuristics (Russell & Norvig, 2010, p. 700). Hence, different decision trees implementations exist (e.g., J48, ADTree, REPTree), which nevertheless keep the core idea the same (Barbon, Igawa, & Bogaz Zarpelão, 2017).

Overall, classification and regression tasks can effectively be treated by decision trees-based approaches (Russell & Norvig, 2010, pp. 697-707). In social media user profiling, decision trees can be used e.g., to classify users upon political orientation based on their profile data, linguistic content, and social network (Chen, Zhu, Guo, & Liu, 2014, p. 163), to detect online hoaxes and frauds (Peng, Detchon, Choo, & Ashman, 2017, p. 2), to predict potential followers of specific companies (Pang, Jiang, & Chen, 2013, p. 398), etc.

Dimensionality reduction

Approaches: *PCA (clustering), PCFA (principal component factor analysis) (clustering), dimensionality reduction, tensor reduction for dimensionality reduction*.

Dimensionality reduction is distinguished as a separate category of approaches for user profiling among others by (Zhang & Bors, 2019, p. 220).

Dimensionality reduction, the second most popular unsupervised learning technique besides clustering, desires to identify the most redundant fields of the input data to keep only the dimensions mostly responsible for the variability, i.e., to reduce the n-dimensional input data to m-dimensional data with $n < m$ keeping the variance at a high level (Cady, 2017, p. 135).

Consider as an example a simple case of reducing the 2-dimensional data with two features height and weight to the 1-dimensional vector by Principal Component Analysis (PCA) approximation as depicted on Figure 9.

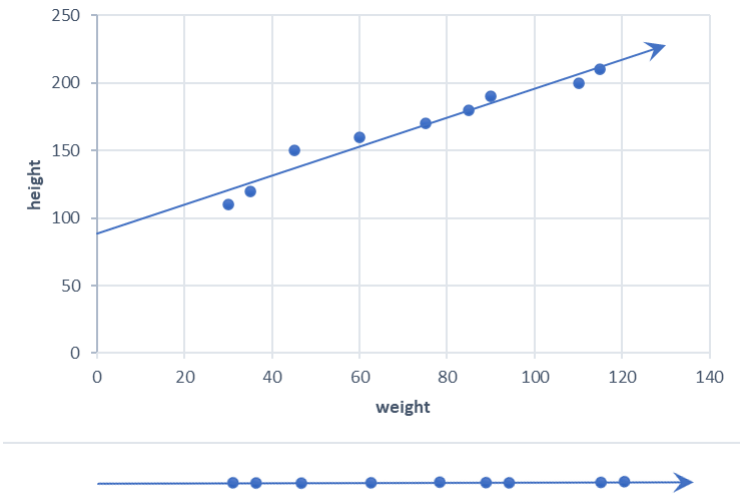

Figure 9. Dimensionality reduction of 2-dimensional data to 1-dimensional vector

In social media user profiling, dimensionality reduction is mentioned mostly as a preliminary step to reduce the dimension of the input data prior to applying other advanced approaches (Zhang & Bors, 2019) (Laere, Buyl, & Nyssen, 2014) (Laere S. , Buyl, Nyssen, & Debruyne, 2017) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016), usually motivated by computational concerns, in particular when processing various multimedia data (Cady, 2017, p. 135).

Ensemble Learning

Approaches: *averaging models, stacking models, dynamic weighted ensemble (DWE), ensemble learning, majority vote, random forest, stacked model to do classifier stacking, stacked SVM, stacking and boosting enhanced ensemble, two ensembles: SVM with RBF kernel, XGBoost, gradient boosting*.

Ensemble Learning is distinguished as a separate category of approaches for user profiling among others by (Russell & Norvig, 2010, p. 748) (Guo, et al., 2016).

The basic idea of ensemble learning is to construct a model with high prediction performance (i.e., strong learner) by combining in a specific way a set of models with low prediction performance (i.e., weak learners) (Chen, Zhang, Chen, Fan, & Gao, 2018, p. 37). One of the best-known and straightforward to understand approaches in this category is random forest, which by combining multiple decision trees provides better overall performance than any single decision tree (Cady, 2017, pp. 103-104).

As an example, Figure 10 provides a schematic overview of a simple majority voting based on five decision trees to determine whether a particular person is either to be considered as fit or as unfit.

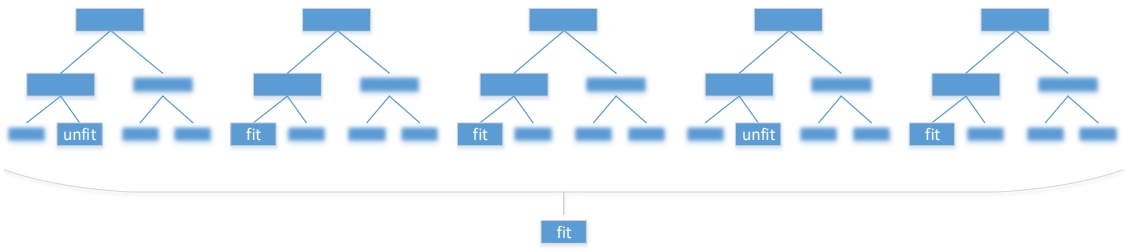

Figure 10. Random forest to decide whether a person is fit or unfit

Very good performance of ensemble-based models on a variety of tasks makes them common approaches of choice in general, and in particular in social media user profiling appliable e.g., to infer Twitter users' nationalities by gradient boosting based approach from language, hashtags, geographical locations, profile pictures, social links (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016, p. 319), to estimate sentiment of tweets by XGBoost with 100 estimators (França, Goya, & Camargo Penteado, 2018), and others.

Graph theory algorithms

Approaches: *association rule mining, standing ovation model (SOM) based on, bird flocking, CENE (network embedding and content), community detection, graph based approaches (centrality, betweenness), graph embedding, graph embedding algorithms (LINE, PUHE), graph embedding learning, graph partitioning, Graph Theoretic Analysis, graph theory, graph-based (session-based temporal graph), graph-based algorithms, heterogenous graph embeddings, large-scale information network embedding (LINE), network analysis (graph-based), network embeddings, normalized graphs, relational graph, social graph*.

Graph theory algorithms are distinguished as a separate category of approaches for user profiling among others by (Kandias, Mitrou, Stavrou, & Gritzalis, 2014).

Graph theory is the study of graphs, i.e., mathematical structures $G = (V, E)$ to model relations between objects through set of vertices $V = V(G)$ and set of edges $E = E(G)$ between these vertices (Drmota, Gittenberger, Karigl, & Panholzer, 2007, p. 58). There is a large variety to incorporate different types of vertices and edges on graphs, resulting in the possibility to represent the most complex models, as exemplarily outline on Figure 11 that contains undirected and directed edges, edges of different weight, vertices of different types belonging to different subsets, etc. (Tang, et al., 2015, p. 1068).

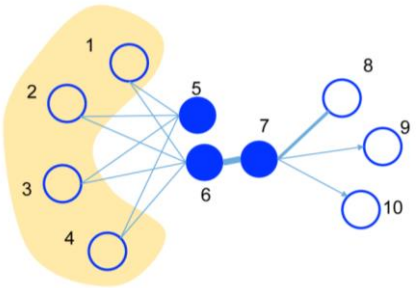

Figure 11. Graph with different vertices and edges types (Tang, et al., 2015, p. 1068)

Consider as an example social graphs that simply reflect users' connections (e.g., with friends, followers, followees, and also to specific groups, etc.) in graph form, after the construction of which a particular graph-embedding algorithm (e.g., LINE) can be applied to obtain the representation of each user in social network required for further evaluation (Xu, et al., 2019, p. 2).

In social media user profiling, the usage of graph-based algorithms is hence natural, justified by the common network structure of social media by users' connections. On the one hand, usually, graph-based approaches are applied in conjunction with other approaches such as to construct a particular social graph to perform further analysis on which (Yin, Thapliya, & Zimmermann, 2018), to obtain network embeddings prior to running an ANN-based algorithm (Zhang, Fu, Jiang, Bao, & Zeng, 2018, p. 12). On the other hand, independent graph-based approaches are also applied e.g., to model users' short-term and long-term interests over time through a specific graph-based structure (Zarrinkalam, Kahani, & Bagheri, 2019, p. 97), etc.

Linear models

Approaches: *linear regression (adapted balance winnow algorithm), linear regression, logistic regression, modified balanced winnow algorithm (learning a linear classifier), multinominal logistic regression (MLR)*.

Linear models are distinguished as a separate category of approaches for user profiling among others by (Russell & Norvig, 2010, p. 717) (Zhang & Bors, 2019).

Linear models represent another type of rather simple yet widely used machine learning approaches (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015; Chen, Zhu, Guo, & Liu, 2014; Zhang, Fu, Jiang, Bao, & Zeng, 2018; Zheng, Li, Zhang, Xie, & Zhong, 2019; Xu, Tadesse, Fei, & Lin, 2019; Buraya, Farseev, & Filchenkov, 2018; Gu, et al., 2018; Guo, et al., 2016; Kandias, Mitrou, Stavrou, & Gritzalis, 2014).

The simplest case of linear models is a univariate linear regression, i.e., linear function on continuous values of the form $y = w_1 x + w_0$, where y is the output value, x is the single input value, and both w_0 and w_1 stand for weights, i.e., the coefficients to be learned. The weights' learning is traditionally done using the squared loss function summed over all the input-output pairs (i.e., supervised learning) (Russell & Norvig, 2010, p. 718). Consider as an example the task to infer weekly social media online time from the number of weekly activities (e.g., posts, shares,

likes, comments, etc.), with input-output pairs plotted as dotes and linear regression respectively as straight line on Figure 12. Hence, should previously unknown user have in average 15 activities per week, then based on constructed linear regression the predicted average weekly social media online time is around 10 hours.

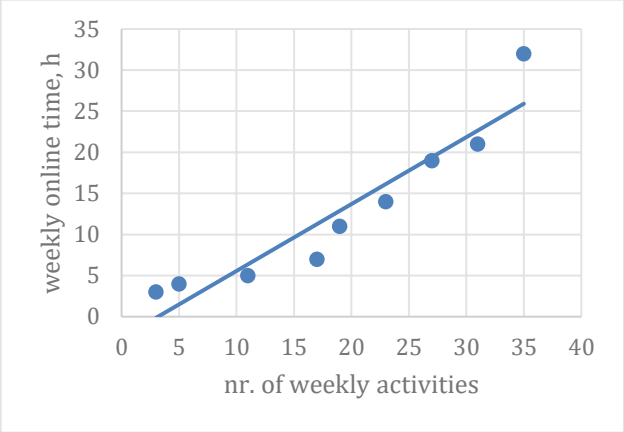

Figure 12. Linear regression to predict weekly online time from weekly activities

One of the straightforward extensions to univariate linear regressions is the class of multivariate linear regressions, that is getting multiple input values instead of only single one (Russell & Norvig, 2010, pp. 720-723). Another, more specific and very common type of linear models are logistic regressions that utilize linear functions to conduct classifications (Russell & Norvig, 2010, pp. 723-727).

Hence, although regression tasks are primarily seen as the main application are of linear models, classification tasks can also be effectively treated by linear models (Russell & Norvig, 2010, pp. 717-727). In social media user profiling, variations of linear models can be e.g., used to infer different traits from user communications based on lexical features (Zheng, Li, Zhang, Xie, & Zhong, 2019, p. 3), to improve user creditworthiness classification based on n-gram and topic features (Guo, et al., 2016), to predict individual psycho-demographic profiles from likes on Facebook (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016, p. 317), etc.

Natural Language Processing

Approaches: *author topic model, bag of concepts, bag-of-words (BOW), CALGARI and KL-divergence scored, continuous bag-of-words (CBOW), corr-LDA, Cross-Media-LDA (CMLDA), frequent pattern mining (FPM) (topic detection), frequent terms (bag of words), geographic topic models, geographical topic models by utilizing statistical topic models, Gibbs sampling in*

location-based topic models, heuristic approaches (TF, TF-IDF, TI-TextRank, FTF), hypertext induced topic search (HITS), labeled-LDA, language models, LDA, L-LDA, mm-LDA, n-grams, N-grams authorship verification (AV), NLP, POS (part of speech), Rocchio, semi-supervised topic model, sentiment analysis (topic models), supervised topic modeling, text classification, text feature + HAN, TF-IDF, TF-IDF features, topic model, topics model, TS-LDA, twitterLDA, unsupervised method for topic detection, unsupervised multilingual approach, weighed-Node2Vec, Word2Vec, Linguist Quantifier driven Tag Determination (LQT), ITT, information filtering (IF), Pattern Taxonomy Model (PTM), PDS, IPE, QBLDA, dynamic user attribute model (DUAM), vector space model (VSM).

Natural Language Processing (NLP) is distinguished as a separate category of approaches for user profiling among others by (Zhang & Bors, 2019) (Russell & Norvig, 2010).

NLP approaches are closely related to the category of probabilistic and statistical models, since natural language models are rather to be defined as a probability distribution over sentences than as a definitive set, and there is also often no single meaning for a sentence, but rather a probability distribution of possible meanings (language ambiguity) (Russell & Norvig, 2010). Thus, some of the most prominent NLP approaches, such LDA or *n*-gram based models, clearly utilize statistical characteristics of the available data as their base elements (Russell & Norvig, 2010, pp. 860-885). Nevertheless, the main defining focus of NLP models is clearly on tasks of processing natural language, such as text classification, information retrieval and extraction, etc. (Russell & Norvig, 2010, p. 860)

Consider some of the simplest and the most straightforward NLP approaches, namely those based on *n*-grams, i.e., sequences of *n* items (e.g., characters or words) from a sample of text (Russell & Norvig, 2010, pp. 860-885). As an example, 1-grams (unigrams), 2-grams (bigrams), 3-grams (trigrams) of the phrase “to be or not to be” are as depicted in Table 11.

Unigrams	Bigrams	Trigrams
to, be, or, not, to, be	to be, be or, or not, not to, to be	to be or, be or not, or not to, not to be

Table 11. 1-, 2-, 3-grams of the phrase "to be or not to be"

Then, based on the distribution of *n*-grams in a specific sample text it is possible to provide next word predictions for a particular input as depicted in Table 12.

Input	Next word predictions
How	many - much - do - does - to - …
How many	days - weeks - people - countries - …
How many days	until - till - in - since - …
How many days until	Halloween - election - …

Table 12. Example of next word predictions

Since language is the major expression and communication mean on social media, hence also NLP is appliable for a large variety of tasks of social media user profiling such as e.g., to determine users that engage in e.g., health-related information sharing (Zhang & Bors, 2019, p. 214), to estimate sentiment of tweets in respect to a specific event (França, Goya, & Camargo Penteado, 2018), to calculate profile similarities of different users (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017, p. 2), and many others.

Nearest neighbour models

Approaches: *CBR (case-based reasoning), collective classification, k-NN, nearest neighbour distribution over ODP (Open Directory Project), neighbourhood-based methods, similarity-based methods, user similarities by k-NN.*

Nearest neighbour (NN) models are distinguished as a separate category of approaches for user profiling among others by (Eke, Norman, Shuib, & Nweke, 2019).

Nearest neighbour models are so-called nonparametric models and are also very important and widely used (Anand & Mampilli, 2014; Valsamis, Psychas, Aisopos, Menychtas, & Varvarigou, 2017; Peng, Detchon, Choo, & Ashman, 2017; Pang, Jiang, & Chen, 2013; Li, et al., 2019; Arain, et al., 2017; Li, Yang, Xu, Wang, & Lin, 2019; C C & Mohan, 2019; Barbon, Igawa, & Bogaz Zarpelão, 2017).

k-nearest neighbours (*k*-NN), the most prominent nearest neighbour model, first proceeds by looking for *k* examples nearest to the new (unlabelled) example of interest in the terms of chosen similarity metrics, and then, depending on the underlying task, either chooses the most common class among these *k* neighbours for the new example (i.e., classification) or computes mean, median, etc. of the *k* neighbours (i.e., regression task) (Russell & Norvig, 2010, p. 738). Consider

as an example the problem to classify users as fit or unfit based on their age, weekly fast-food consumption and number of days exercising per week. The exemplarily data, plotted on Figure 13, suggests that a new user to be classified who consumes fast-food and exercises on 3 days a week is most probably unfit if under 30 and fit if over 30.

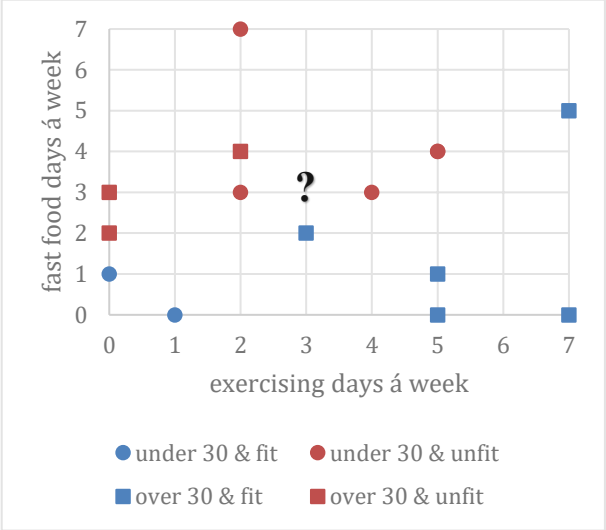

Figure 13. k-NN to decide whether a person is fit or unfit

The computational complexity of *k*-NN respectively grows with the number of available instances, since these are required for the classification of new instances of interest (i.e., instance-based learning). In other words, *k*-NN cannot be characterized by a bounded set of parameters and thus nonparametric model, i.e., in contrast to e.g., decision trees or linear models (Russell & Norvig, 2010, p. 737).

In social media user profiling, different nearest neighbour approaches can be used e.g., to predict the rating a user would give to a particular item (e.g., movie or a TV program) based on that item's ratings by similar users (Anand & Mampilli, 2014, p. 2430; Valsamis, Psychas, Aisopos, Menychtas, & Varvarigou, 2017, p. 2), user profile authorship attribution (i.e., to find different profiles that belong to the same person) (Peng, Detchon, Choo, & Ashman, 2017), to identify companies' potential customers (Pang, Jiang, & Chen, 2013, p. 398), etc.

Probabilistic and statistical models

Approaches: *affiliation graph model (network structure, probabilistic graphical model), analysis of variance (ANOVA), asynchronous stochastic gradient algorithm (ASGD), Bayesian classification, Bayesian inference, Bayesian networks, Bayesian networks and ontologies,*

Bayesian personalized ranking, Bayesian technique, belief function reasoning, Chinese restaurant process (statistics/probability), collective naïve Bayes, composite Gaussian Process (GP), Gaussian distribution, Gaussian Mixture Model, Gaussian relational topic model, graphical models, Hidden Markov Models (HMM), hierarchical Bayesian model, HMRF (Hidden Markov Random Field), HMRF-KMEANS, incremental Bayesian online updates, Markov chains, Markov logic network (MLN), Markov random field, Maximum Likelihood Estimation, MCMC, modded SVD (modSVD), MRF (Markov Random Field), Naïve Bayes, naïve Bayes classifier, naïve Bayes multinominal (NBM), outliers determined by interquartile ranges (IQR), probabilistic approaches (Explicit semantic analysis (ESA)), probabilistic framework, probabilistic inference (for user location), probabilistic latent semantic analysis (PLSA), probabilistic matrix factorization (PMF), probabilistic model, probabilistic topic model, probability distributions, probability models, relational naïve Bayes classifier, Restricted Boltzmann Machines (RBMs), SALSA (stochastic approach for link-structure analysis), singular value decomposition (SVD), statistical analysis, statistical classifier, statistical modeling, stochastic gradient descent classifier, stochastic topic model, TimeSVD, transferable belief model (TBM), TrustSVD (latent factor model), unified discriminative influence model, unified discriminative influence probabilistic model, discriminative influence model, generative influence models, generative relationship influence models, dependence distributions, factor graph model, SoRec (social regularization).

Probabilistic and statistical models are distinguished as a separate category of approaches for user profiling among others by (Eke, Norman, Shuib, & Nweke, 2019).

The category of probabilistic and statistical models covers some of the highest number of different approaches, which are in turn interrelated with other categories, since probability and statistics constitute some of the fundamental areas for most of the approaches of other categories (Russell & Norvig, 2010, pp. 7-9). Nevertheless, because some approaches are much closer to their underlying probabilistic and statistical concepts than the to the other specific categories of social media user profiling, thus they are combined into this separate category of probabilistic and statistical models.

Outcomes of different models usually need to be analysed for statistical importance (Russell & Norvig, 2010, p. 25). Furthermore, partial observability, nondeterminism, or a combination of the two result in the necessity to handle uncertainty, hence using probability theory and statistical modelling (Russell & Norvig, 2010, pp. 480-688), as e.g., most of the modern approaches to uncertain reasoning build on the Bayes' rule (Russell & Norvig, 2010, p. 9). Consider as example

the data provided in Table 13 to determine person's fitness, and the assumption of conditional dependence.

exercising days nr. á week	fast food days á week	over 30?	fit?
0	1	no	yes
4	3	no	no
5	4	no	no
1	0	no	yes
3	5	no	no
2	3	no	no
5	4	no	no
7	3	yes	yes
0	2	yes	no
0	3	yes	no
2	4	yes	no
3	2	yes	yes
5	0	yes	yes
7	0	yes	yes
5	1	yes	yes

Table 13. Sample data of fitness level of different people

Should be known that a new person exercises 3 times à week, consumes fast food 3 times á week, and is not over 30 years old, then based on calculating the probabilities for that particular person being fit or unfit using Bayes' rule it can be concluded that the chances are higher this person is unfit. The computation of the probabilities is as follows:

$$P(exercising = 3\ AND\ fastfood = 3\ AND\ over30 = no \mid fit) * P(fit)$$
$$= 1/7 * 1/7 * 2/7 * 7/15 = 0.0027$$
$$P(exercising = 3\ AND\ fastfood = 3\ AND\ over30 = no \mid unfit) * P(unfit)$$
$$= 1/8 * 3/8 * 5 * 8 * 8/15 = 0.0156$$

Probabilistic and statistical modelling approaches are found to be appliable to social media user profiling in particular to derive user's gender, marital status, and other attributes from such information as groups membership, likes, views, etc. (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015), to predict user's Big Five personality traits from Twitter data (Li, et al., 2019, pp. 274-275), and many others.

Social recommender systems

Approaches: *autoencoder-based social recommender system (AESR), binary relevance (BR) (user tags prediction), classic rank (CLR), classifier chains (CC), co-factorization machines (CoFM), collaborative filtering (activity recommendation engine), Collaborative Filtering (memory-based CF, matrix factorization (SVD, LDA, ALS)), collaborative filtering rank (CFR), content-based and graph-based features, content-based and preference-based filtering, content-based collaborative filtering, content-based systems, content-based user tag recommendation, context-aware recommender systems (CARS), dimensionality reduction through network embedding paradigm (matrix factorization), FOAF, FREQ (frequency of tags), friends-based collaborative filtering, fuzzy logic (OWA (ordered weighted averaging) operators), generalist recommender system kernel (GRSK), generalized matrix factorization, graph-based user tag recommendation, individual filtering (user preferences), item-based CF, latent factor model, LTPA (local tag propagation), matrix decomposition techniques (specifically non-negative matrix factorization (NMF)), matrix factorization, memory-based CF, model-based CF, model-based recommendation (matrix factorization, probabilistic latent factor models), most popular friends (MPF), neighborhood-based CF, neural recommendations (neural networks and collaborative filtering), non-negative matrix factorization (NMF), popularity rank (PR), rating-based systems, recommender systems (social recommendation), social pertinent walker (SPTW), social tagging system (STS), social-based collaborative filtering (CF), social-based filtering (friends' preferences), temporal and social probabilistic matrix factorization, temporal influence correlations (TIC), tensor factorization (TF) models, TopicMF (matrix factorization), user-based CF, utility based user profiling mining (UUPM), Crisp User Profile based Recommendations (CUP), Rank based Degree of Feature (RDF), EIUCF, QICE, injected preferences fusion (IPF), demographic systems, entropy-based model (EBM), utility-based systems (UB)*.

Social recommender systems (SRS) are distinguished as a separate category of approaches for user profiling among others by (Zhou, Xu, Li, Josang, & Cox, 2012).

The defining characteristic of the SRS is to incorporate social media data into recommender systems (C C & Mohan, 2019). Approaches of recommender systems are information filtering tools to provide the most relevant and accurate content to users of a particular service based on the users' behaviour, i.e., to support the users in discovering personalized information of interest from large amounts of available complex and dynamic information (C C & Mohan, 2019, p. 1937). Furthermore, approaches of recommender systems are usually divided into specific categories such as collaborative filtering, content-based filtering, hybrid systems, etc. (C C & Mohan, 2019, p. 1937) (Eke, Norman, Shuib, & Nweke, 2019, pp. 144916-144917). Figure 14 contains a very simplistic outline comparing both collaborative and content-based filtering of recommender systems.

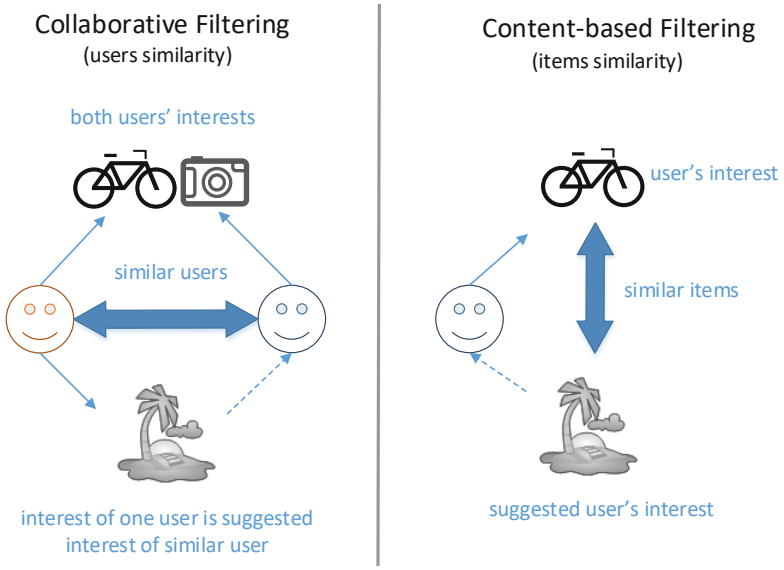

Figure 14. Collaborative and content-based filtering examples

Recommender systems are overall very popular in the social media domain, newly also in social media user profiling, e.g., to suggest points of interest (Wang, Zhong, Yang, & Jing, 2018), tourism destination (Zheng, Luo, Sun, Zhang, & Chen, 2018), to inference social media users' location, occupation, education, religion (Yang, Xiao, Tong, Zhang, & Wang, 2015), etc.

Social semantic web

Approaches: *association use mining, Business Semantics Management, collaborative ontology engineering methods, combine semantic context and social network information, Dublin Core, expert systems, explicit semantic analysis, filtering techniques, fitness buddies recommendation engine, GOSPL, GOSPL with D2RQ, hierarchical interest graph (from Wikipedia category*

graph), knowledge graphs, knowledge-based systems (KB), Latent Semantic Analysis, Latent Semantic Analysis (LSA) using matrix factorization technique, latent semantic hashing, map image to knowledge graph entities, ontologies, ontology based, ontology engineering project, ontology to categorize results, ontology-based recommendations, ontology-based user models, OpenDNS, DBpedia, OWL, RDF, semantic methods to recommend friends, semantic relationships, semantic structures, semantic technologies for interlinking social websites, semantic trees, semantically enrich user profiles by using association rules, spreading activation algorithm (e.g., algorithm over semantic networks), syntactic and semantic algorithms, user hierarchical knowledge graphs, user ontology profiling, DILIGENT, HCOME, SIOC.

Social semantic web (SSW) is distinguished as a separate category of approaches for user profiling among others by (Orlandi, 2012) (Eke, Norman, Shuib, & Nweke, 2019).

Semantic web intends to improve web by achieving computer-readability, i.e., to recognize and to infer information on web by assigning semantic to it (Bok, Yoon, & Yoo, 2019, p. 28682). Social semantic web is simply a combination of semantic web and social web to find additional meaningful knowledge in web data by utilizing human relationships and interaction in semantic web technology (Bok, Yoon, & Yoo, 2019, p. 28682). The key element of the semantic web are ontologies, which hold formal descriptions and specifications of concepts (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013, p. 169). As an example, Figure 15 contains schema of the researcher profile proposed by (Tang, Yao, Zhang, & Zhang, 2010, pp. 5-6) constructed by extending the FOAF ontology.

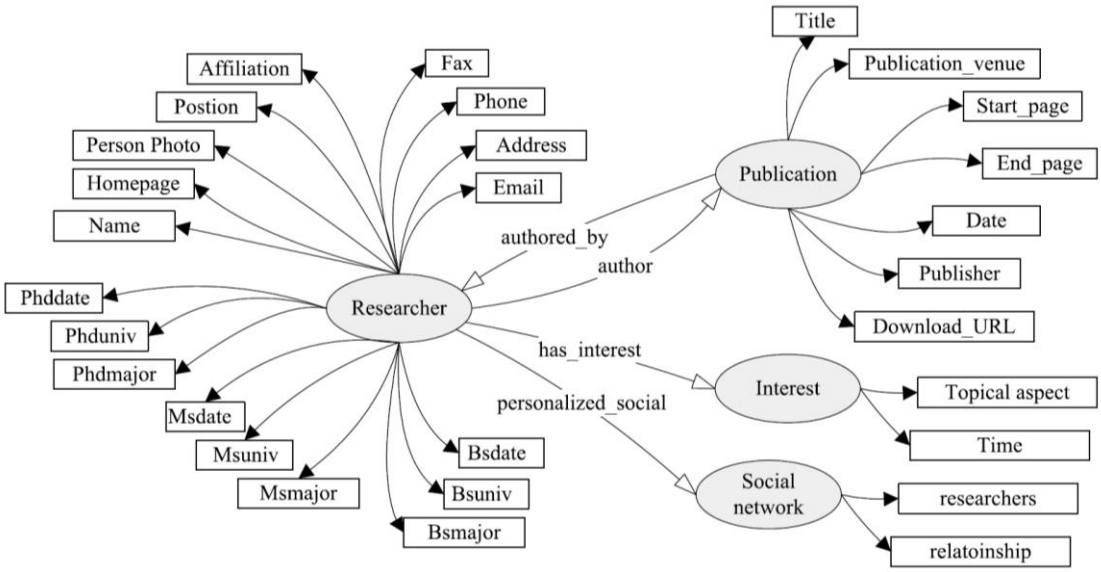

Figure 15. Schema of the researcher profile by extending the FOAF ontology (Tang, Yao, Zhang, & Zhang, 2010, pp. 5-6)

Semantic web can play an important role in social media user profiling, as e.g., to define and model users' profiles through extending semantic ontologies (Tang, Yao, Zhang, & Zhang, 2010, pp. 5-6) (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013) to automatically populate user profile ontology with assertions of interests and intentions as next (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013), etc.

Support vector machines

Approaches: *latent SVM (LSVM), non-linear (radial basis function) kernel SVM, SVM, SVM classifier, SVM with linear kernel (linear kernel SVM, linear SVM, SVM with linear kernel)*.
Support vector machines (SVM) are distinguished as a separate category of approaches for user profiling among others by (Eke, Norman, Shuib, & Nweke, 2019).

Support vector machine (SVM) is often seen as an excellent ML approach to try as first, hence widely adopted in a variety of use cases (Ma, et al., 2015; Chen, Zhu, Guo, & Liu, 2014; Tang, Yao, Zhang, & Zhang, 2010; Chen, Zhang, Chen, Fan, & Gao, 2018; Pipanmaekaporn & Kamonsantiroj, 2015; Zhang, Fu, Jiang, Bao, & Zeng, 2018; Zheng, Li, Zhang, Xie, & Zhong, 2019; Zhuang, Ma, & Yoshikawa, 2017; Gu, et al., 2018; Peng, Detchon, Choo, & Ashman, 2017) (Guo, et al., 2016; Pang, Jiang, & Chen, 2013; Li, et al., 2019; Kandias, Mitrou, Stavrou, & Gritzalis, 2014; Lee, Hussain, Rivera, & Isroilov, 2018; Dougnon R. , Fournier-Viger, Lin, &

Nkambou, 2016; Fang, Sang, Xu, & Hossain, 2015; Hoang & Lim, 2017; Li, Yang, Xu, Wang, & Lin, 2019; Barbon, Igawa, & Bogaz Zarpelão, 2017).

By using SVM the main aim is to construct a maximum margin separator (support vectors), i.e., a decision boundary in form of a so-called hyperplane with the largest possible distance to available input items (Russell & Norvig, 2010, p. 744). Although linear separating hyperplanes form the basis of SVM, there is also the possibility to utilize higher dimensions for non-linearly separable input data by using so-called kernel trick (e.g., polynomial or radial basis function kernels) (Russell & Norvig, 2010, p. 744). Consider as an example data on Figure 16 with respective linear hyperplane.

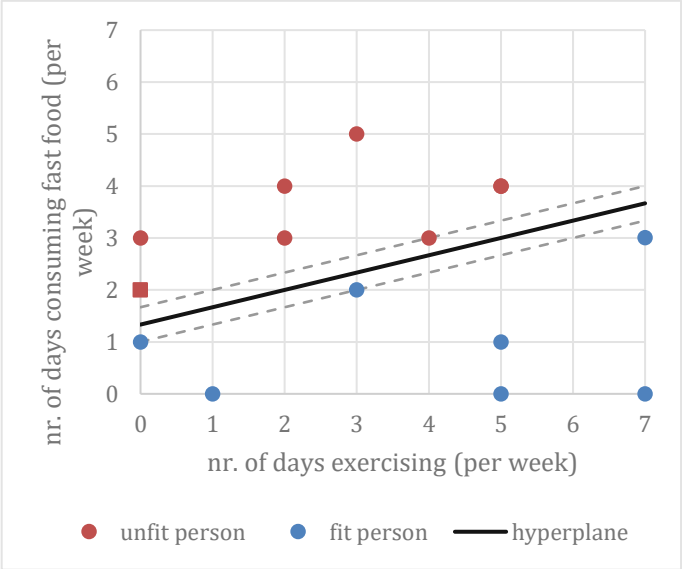

Figure 16. SVM example of hyperplane separating fit from unfit people

Since SVM use input-output pairs (i.e., supervised learning) and potentially retain the complete learning data (i.e., instance-based learning), hence SVM is also a nonparametric method similar to k-NN that cannot be characterized by a bounded set of parameters (Russell & Norvig, 2010, p. 744). In the praxis, however, only a certain portion of data is retained after learning, often small constant times the number of constructed hyperplane dimensions, thus combining the advantages of parametric and nonparametric methods (Russell & Norvig, 2010, p. 744).

SVM are mainly used for classification problems, although regression tasks can also effectively be approached with SVM (Russell & Norvig, 2010, pp. 744-748). In social media user profiling, SVM can be used e.g., to classify user's gender, age, regional origin, and political orientation from tweets (Chen, Zhu, Guo, & Liu, 2014, p. 163; Dougnon R. , Fournier-Viger, Lin, &

Nkambou, 2016, p. 319; Faralli, Stilo, & Velardi, 2015, p. 3), to predict the profession of Twitter users (Chen, Zhang, Chen, Fan, & Gao, 2018, p. 31), to determine text authorship and to detect hoaxes and frauds (Peng, Detchon, Choo, & Ashman, 2017).

6.2.3 Categories of explainability techniques for social media user profiling

The extracted terms of explainability approaches for social media user profiling in credit scoring, as identified in the subsection 5.3., are categorized by following a hybrid approach (Usman, 2015, p. 124) that combines traditional top-down and bottom-up approaches (Broughton, 2015) in the following way: the initial categories of explainability approaches for social media user profiling in credit scoring are extracted and terminology controlled from the selected publications of the SLR on explainability techniques for social media user profiling in credit scoring in the subsection 4.3. (i.e., top-down); identified single explainability techniques for social media user profiling in credit scoring are successively assigned to matching categories, creating new category in case no matching category for a particular term exists (i.e., bottom-up).

During the categorization procedure, the decision was made not to distinguish scope (global vs. local) (Singh, Sengupta, & Lakshminarayanan, 2020, pp. 2-3) (Carvalho, Pereira, & Cardoso, 2019, pp. 14-15), model-specificity (model-agnostic vs. model-specific) (Singh, Sengupta, & Lakshminarayanan, 2020, p. 2) (Carvalho, Pereira, & Cardoso, 2019, pp. 12-13), and timing of explanation generation (pre-model vs. in-model vs. post-hoc) (Singh, Sengupta, & Lakshminarayanan, 2020, p. 3) (Carvalho, Pereira, & Cardoso, 2019, p. 12) (Preece, 2018, p. 67) properties of explainability techniques into separate higher-level categories, since descriptions of extracted explainability techniques often lack clear information in regard to these properties. Where such information is provided, the respective note is also added in the descriptions below. As the result, the identified categories for the explainability techniques for social media user profiling in credit scoring are as follows:

- Decision Tree based explanations
- Deep explanations
- Explainable surrogate models
- Features importance
- Model combination
- Prototype selection
- Recommender systems explanations
- Rules based explanations

- Salient masks
- Sensitivity analysis
- Textual justification
- Transparent model types
- Visual techniques

These categories are described in the following subsections in details, providing the information on which of the extracted terms belong to which of the identified categories.

Decision Tree based explanations

Techniques: *combination of genetic algorithms with decision trees or rules, Confident Decision Tree (CDT), Decision Diagrams, decision trees, Single Tree, Single Tree Approximation (STA), Simplified Tree Ensamble Learner (STEL), Tree Metrics, tree regularization, Tree Space Prototype (TSP), TreeView, DecText, oblique tree sparse additive models (OT-SpAMs), PALM, inTrees, GPDT, Trepan, tsp*.

Decision Tree based explanations are distinguished as a separate category of explainability techniques among others by (Guidotti, et al., 2019, p. 17).

The transparent nature of decision trees is perfectly suited to serve as an explanator for other models (Rosenfeld & Richardson, 2019). Tracing upwards from the predictions in leaf nodes through the splits yields to understanding which features and what cut-off values were used in producing a particular prediction, hence also what attributes are the most important, etc. (Rosenfeld & Richardson, 2019, p. 10) (Carvalho, Pereira, & Cardoso, 2019, p. 14). At the same time, very large trees are much harder to follow (Rosenfeld & Richardson, 2019, p. 10), and features of models that use highly engineered, anonymous, or opaque features are losing their explicitness (Carvalho, Pereira, & Cardoso, 2019, p. 14).

Models to be explained by decision trees first need to be accordingly approximated, with one of the most prominent such techniques called Single Tree Approximation (STA) (Guidotti, et al., 2019, p. 17). By STA, a large set of instances is submitted to the model to be explained, subsequently analysing the different predictions and constructing the respective decision tree that is explainable by its nature (Guidotti, et al., 2019, p. 23).

Decision Tree based explanations are predominately post-hoc techniques (Guidotti, et al., 2019), majority being model-specific with some adoptable as model-agnostic explanators (Guidotti, et al., 2019), primarily for global, but also for local explanations (Guidotti, et al., 2019, p. 17).

Deep explanations

Techniques: *Deep Attention-Based Representations for Explanation/Explainable Generative Adversarial Networks (DARE/X-GANS), deep explanation, Deep Hierarchical Generative Models, Deep SHapley Additive exPlanations, deep Taylor decomposition, DeepTaylor, deep tensor networks, DeepDreams, DeepExplain, DeepTune, backpropagation of the gradients, back-propagation-based methods, guided backpropagation, capsule network, DeConvNet, deconvolution, excitable network attractors (ENAs), GNNExplainer, GroupINN, interpretable convolutional neural networks, network propagation technique based on deconvolutions to reconstruct input image patterns that are linked to a particular feature map activation or prediction, MDNet, Rational explanations, Reflexive explanations, shallow models, uniform probabilistic framework*.

Deep explanations are distinguished as a separate category of explainability techniques among others by (Gunning & Aha, 2019, p. 45).

The idea behind deep explanations is to adapt deep learning approaches to learn more explainable features, more explainable model representations, or to directly utilize them in generating explanations (Gunning & Aha, 2019, p. 45). Architecture- or domain-specific techniques together with, more standard, attribution-based techniques constitute the most common deep explanation types, i.e., either developing methodology and validating it on a particular problem or performing a separate analysis based on pre-existing techniques that assign an attribution value (in other words, relevance or contribution) to each input of a deep neural network (Singh, Sengupta, & Lakshminarayanan, 2020, pp. 4-12).

Deconvolution, a specific backpropagation-based technique, hence also attribution-based (Singh, Sengupta, & Lakshminarayanan, 2020, pp. 4-6) (Zihni, et al., 2020, p. 6), proceeds e.g., by developing to a convolution neural network (CNN) that is required to be explained an additional CNN running on the output to undo the operations of the original CNN (Rio-Torto, Fernandes, & Teixeira, 2020, p. 374).

Deep explanations are applied either as in-model or post-hoc (Singh, Sengupta, & Lakshminarayanan, 2020, pp. 4-12) (Gunning & Aha, 2019, p. 45), mostly model-specific (Carvalho, Pereira, & Cardoso, 2019, p. 20), and predominately global in scope (Singh, Sengupta, & Lakshminarayanan, 2020, pp. 4-12).

Explainable surrogate models

Techniques: *approximate an interpretable model for the black-box model, interpretable model extraction, mimic learning, Local Interpretable Model-Agnostic Explanation (LIME), post hoc application of supervised learning with support vector machines, surrogate model, SP-LIME, probabilistic generative model, proxy models, reducing complex NN*.

Explainable surrogate models are distinguished as a separate category of explainability techniques among others by (Carvalho, Pereira, & Cardoso, 2019, p. 13).

Explainability techniques that utilize the idea of surrogate models aim to generate an intrinsically explainable, e.g., explainable by design, surrogate model that approximates the model desired to be explained (Carvalho, Pereira, & Cardoso, 2019, p. 13).

LIME (Local Interpretable Model-Agnostic Explanation) is a very popular explainability technique overall and of the category of explainable surrogate models in particular. LIME conducts learning of a simple, interpretable model based on perturbed inputs and responses to them from a more complex model that is to be explained (Preece, 2018, p. 66). Then, in case of explaining an image classification decision, contiguous regions of pixels (so-called super-pixels) that similarly contribute towards the decision in favour of a particular class are highlighted to the user as an explanation (Preece, 2018, p. 66).

Explainability techniques based on explainable surrogate models are, clearly, post-hoc techniques (Carvalho, Pereira, & Cardoso, 2019), mostly model-agnostic (Guidotti, et al., 2019, pp. 28-29), and suitable for both global or local explanations (Carvalho, Pereira, & Cardoso, 2019, p. 13).

Features importance

Techniques: *Accumulated Local Effects Plot, anchors, attribute explanations, coalitional game theory based, Deep Learning Important FeaTures (DeepLIFT), expert-determined features relevance, explanation in terms of input variables, feature analysis, feature attribution, feature*

engineering, feature importance, feature influence, feature summary, FINE (feature importance in nonlinear embeddings), feature extraction and explanation extraction framework, Feature Interaction, feature sentiments, interaction and feature importances, One-variable-at-a-Time approach, auditing, BreakDown, CFS, e-LRP, Layer wise relevance propagation (LRP), z-LRP, filters, GoldenEye, individual conditional expectation, investigation of deep representations, multivariate filters, SHapley Additive exPlanations (SHAP) values.

Features importance is distinguished as a separate category of explainability techniques among others by (Guidotti, et al., 2019, p. 17).

The idea behind explainability techniques by feature importance is to provide the set of features together with their respective weights (i.e., importance) for the output (Guidotti, et al., 2019, p. 17). This can be done, for example, based on the theoretic information criterion that estimates the entropy of the model's prediction change considering feature perturbation (Bikmukhametov & Jäschke, 2020, p. 9).

One very prominent feature importance technique was introduced from the coalition game theory, namely Shapley value sampling, which proceeds by computing approximate Shapley values through taking each input feature for a sample number of times, resulting in the description fair distribution of the gains and losses among the input features (Singh, Sengupta, & Lakshminarayanan, 2020, p. 5).

Explainability techniques based on feature importance analysis are usually post-hoc (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019, pp. 22076-22077), model-agnostic explainability techniques (Guidotti, et al., 2019, pp. 28-29), and can be either global or local in scope (Guidotti, et al., 2019, p. 17).

Model combination

Techniques: *combine physics-based models with machine learning, CENTAUR, Combined Multiple Model (CCM), dual neural network system, dual system approach, in-model, in-model joint architecture from explainer and classifier, integration of feature interaction and tree interpretation functionalities into Random Forest program code, knowledge distillation, knowledge extraction, clustering methods, co-clustering approach to gain explainability in a user-item bipartite network, cross-cluster model-log alignment for identifying differences between clusters, combination with other machine learning methods, k-means, kNN, linear*

models, linear dimensionality reduction, LSTM, combine model compression with dimension reduction, Generalized Additive Models (GAMs), nonlinear dimensionality reduction, Principal Component Analysis, second deep network that generates explanations, two different networks to visualize predictions of different network layers, SVM margin, SVM+Prototypes (SVM+P), random sampling, contextual decomposition explanation penalization, explainable question answering system (EQUAS), self-training Grey-Box model, tractable probabilistic logic models (TPLMs), Acceptance Testing, auxiliary criteria, auxiliary data, auxiliary network.

Model combination is distinguished as a separate category of explainability techniques among others by (Guidotti, et al., 2019, p. 24).

Explainability techniques based on combining multiple models proceed by adding a separate model responsible specifically for the explainability in addition to model, the results of which require to be explained (Rio-Torto, Fernandes, & Teixeira, 2020) (Preece, 2018).

An example of the model combination technique is a dual neural network system that consists of image classification model and model to generate textual explanations for the made decision, both combined as sub-systems in one system (Preece, 2018, p. 67).

The nature of explainability techniques based on model combination justifies their primarily application as in-model (Carvalho, Pereira, & Cardoso, 2019, p. 13), model-specific techniques (Carvalho, Pereira, & Cardoso, 2019, p. 13).

Prototype selection

Techniques: *analysis of layers of a 3D-CNN using Gaussian mixture model (GMM) and binary encoding of training and test images based on their GMM components for returning similar 3D images, case-based reasoning, example-based explanation, prototypes, prototype and criticism generation, prototype generation by greedy approach, prototype generation by LP relaxation with randomized rounding, Prototype Selection (PS), Bayesian Case Model (BCM), Compound Critiques, counterfactual explanations, data points, data-dependent, dataset-level, explanation-based learning, genetic programming, GMM and atlas, monotonic constraints, Triplet loss, triplet-loss and k nearest neighbors (kNN) search-based learning strategy, Influence Functions, influential data points, query evidence, root causes of misclassifications, root causes of process model differences, 3-Level Explanation, Bayesian Teaching, similar images, similarity analysis techniques*.

Prototype selection is distinguished as a separate category of explainability techniques among others by (Guidotti, et al., 2019, p. 17).

The goal of prototype selection explainability techniques is to provide an example very similar to the input that led to a specific outcome (Guidotti, et al., 2019, p. 17). Hence, a prototype is representative of a set of similar inputs and is part of observed data or artificially generated to possess particular characteristics (Guidotti, et al., 2019, p. 17).

Bayesian Case Model (BCM), an example of explainability techniques based on prototype selection, learns prototypes as representative samples of specific data clusters, identified on the preceding step (Guidotti, et al., 2019, p. 36). Additionally, BCM learns the so-called subspaces, i.e., sets of features important in identifying each particular prototype (Guidotti, et al., 2019, p. 36).

Prototype selection techniques are mostly post-hoc (Guidotti, et al., 2019), appliable as for local and global scopes (Guidotti, et al., 2019, pp. 35-36).

Recommender systems explanations

Techniques: *Aspect-based Matrix Factorization model (AMF), collaborative filtering method using a tensor modeled by considering the 5Ws with explanations based on template, collaborative-based explanations, community tags to explain recommendations, explain the recommendation process, explainable recommendation, Explainable Matrix Factorization (EMF), explanations for a hybrid recommender system, explanations in time-series recommendation, Factorized Latent Aspect ModEl (FLAME) combining collaborative filtering and opinion mining, graph-based recommendation approach, hybrid approach using collaborative and content-based filtering techniques, integrate explanations into Matrix Factorization, justification explanations, justify why the recommendation might be good for a user, keywords and neighbours and ratings, keywords or user-tags based explanations, leverage topic models to discover explainable latent factors in matrix factorization, MoviExplain, nearest neighbors, neighbor ratings, neighborhood of an instance, neighborhood technique based on cosine similarity, neighbourhood based Collaborative Filtering (CF), neighbourhood style explanation, reviews with ratings to enhance the explainability of matrix factorization, semantic distance (LDSD) algorithm and DBpedia based recommendation system, semantic meaningfulness constraints, semantic monotonicity constraints, semantic property values to*

explain recommendations, semantic web based, SemAuto, SemRec, topic-based explainable recommendation, social collaborative viewpoint regression, social explanations for recommender systems, tag-based explaining approach in graph-based recommender, tags and ratings in a social tagging system with PARAFAC, Tagsplanations, content-based explanations, HFT, CTR, RMR, RBLT, ITLFM, ERBM, explicit factor model, four-order tensor to model users, HIN technique, items, knowledge-based explanations, linked data, MMALFM, Preference-based Organization (Pref-ORG), relational connecting paths, RippleNet, sentiment-based explainable recommendation, sentiment-based tradeoff-oriented explanation approach, separate engine for generating explanations in recommender systems, shared tradeoff properties of a group of products in terms of both static specifications and feature sentiments, shared tradeoff properties of a group of products relative to the top recommendation, structured knowledge bases, TasteWeights, TempEx-Dry, TempEx-Fluid, tradeoff-oriented explanations, tripartite graph encoding user-item-aspect relationships for a review-aware recommendation, TriRank, user-item relevance scores using matrix factorization techniques, users' sentiments on specific aspects, users' sentiments on specific features.

Recommender systems explanations are distinguished as a separate category of explainability techniques among others by (Alshammari, Nasraoui, & Sanders, 2019) (Hong, Akerkar, & Jung, 2019) (Bharadhwaj & Joshi, 2018) (Chen, Yan, & Wang, 2019) (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019).

Explainable recommendation provides explanations for recommended items that users might be interested in, with explanations in the most different forms such as content-based, social, etc. (Hong, Akerkar, & Jung, 2019, pp. 97-98).

Consider specifically numerous examples of explanations in recommender systems that utilize knowledge graphs for recommendation justification: from custom structured knowledge bases about users and items to knowledge graphs automatically retrieved from the semantic web (Alshammari, Nasraoui, & Sanders, 2019, pp. 110565-110566). Information extraction from other (than knowledge graphs) additional data sources is also key to recommender systems explanations of many other explainability techniques as well (Hong, Akerkar, & Jung, 2019, p. 98).

Rules based explanations

Techniques: *Bayesian Rule Lists (BRL, decision lists), Conj Rules, constraint programming for converting linear SVM (and other hyperplane-based linear classifiers) into a set of non*

overlapping and interpretable rules, CPAR (Classification based on Predictive Association Rules), decision rules, decompositional rule extraction, G-REX (decision rules), pedagogical rule extraction, rule extraction, Rule Set, Rule Based Explanator, rule-based methods, rule-based methods for recommender systems, rule-based segmentation, rule-based segmentation followed by a perturbation analysis, decision rules, Two-Level Boolean Rules (TLBR), 1Rule, FRL, REFNE, RxREN, Model Explanation System (MES), Interpretable Decision Sets (IDS), inductive logic programming, search for explanations of clusters of process instances (SECPI), MYCIN, NEOMYCIN.

Rules based explanations is distinguished as a separate category of explainability techniques among others by (Guidotti, et al., 2019, p. 17).

Explainability techniques based on rules provide respective explanations in form of rules of different types, such as decision rules (map observation to action), classification rules (association rules resulting in class label), if-then rules (combination of conditions on input variables followed by specific outcome), m-of-n-rules (m conditions of set of n conditions being verifiable leads to rule considered to be true), etc. (Guidotti, et al., 2019, p. 8).

The explainability technique called RxREN (Rule extraction by Reverse Engineering the Neural networks) proceeds e.g., by first identifying the data range necessary to classify test instance of interest as of a specific class through pruning the insignificant input, and then generates classification rules for each class label, exploiting previously identified data ranges, using reverse engineering (Guidotti, et al., 2019, p. 26).

Rule based explanations are usually post-hoc and model-specific (Guidotti, et al., 2019), capable to be applied for local and global explanations (Guidotti, et al., 2019, p. 17).

Salient masks

Techniques: *complementary examples, Pattern Attribution, PatternNet, saliency, saliency heatmaps (saliency heatmaps, heatmaps of salient regions), saliency maps, saliency masks, salient (highest weighted or most predictive) text features or fragments, salient examples, salient part of the images, salient sentences from text documents using loss gradient magnitudes, salient structures within images related to a specific class by computing the corresponding prediction score derivative with respect to the input image, attributions (saliency maps), CAM, GSInquire, iNNvestigate, rationales as part of the learning process*.

Salient mask is distinguished as a separate category of explainability techniques among others by (Guidotti, et al., 2019, p. 17) (Preece, 2018, p. 67).

The so-called masking principle, i.e., highlighting the determining aspects of the analyzed input regarding produced output, constitutes the main idea behind the salient mask explainability techniques (Guidotti, et al., 2019, p. 17).

Explanations in form of salient examples in text processing domain could have a form of e.g., coherent phrases from the input text that are sufficient to trigger the same prediction as the complete input (Preece, 2018, p. 66).

Salient masks usually belong to post-hoc explainability techniques (Preece, 2018, p. 67), with local or global scope (Rosenfeld & Richardson, 2019).

Sensitivity analysis

Techniques: *sensitivity analysis, sensitivity analysis maps, sensitivity to local variation of the input image, Concept Activation Vectors (TCAV), concept vectors, concept-based explanations, conceptual clustering, TCAV extension by Regression Concept Vectors (RCV), TCAV extension by Uniform unit Ball surface Sampling (UBS), auditing, causal explanations, causal models to explain learning (CAMEL) approach, CAV, Variable Effect Characteristic curve (VEC), Gaussian Process Classification (GDP), UBS, occlusion, perturbation-based explanation, vectors for localized interpretations, Quantitative Input Influence (QII), XRL Interaction (explainable reinforcement learning)*.

Sensitivity analysis is distinguished as a separate category of explainability techniques among others by (Guidotti, et al., 2019, p. 17).

The goal of sensitivity analysis is to evaluate the uncertainty of the outcome of the model to be explained with respect to the different sources of uncertainty in the inputs to this model (Guidotti, et al., 2019, p. 17). Some of the sensitivity measures used in sensitivity analysis are the range, gradient, variance of the prediction (Guidotti, et al., 2019, p. 32).

As an example, sensitivity analysis can be conducted through maximizing the activation of the target neuron in neural networks by performing gradient ascent (finding steepest slope of a function) (Singh, Sengupta, & Lakshminarayanan, 2020, p. 9).

Explainability techniques based on sensitivity analysis are commonly post-hoc (Roscher, Bohn, Duarte, & Garcke, 2020, p. 5), can be model-agnostic (Guidotti, et al., 2019, pp. 31-32) or model-specific (Singh, Sengupta, & Lakshminarayanan, 2020, p. 9), and mostly of local scope (Guidotti, et al., 2019, pp. 31-32).

Textual justification

Model types: *Narrative Generation; natural language caption generation; NL explanations; text descriptions for pictures; textual templates for pre-defined explanations; diagnostic sentence; Argumentation and Pedagogy; argumentation theory based*.

Textual justification is distinguished as a separate category of explainability techniques among others by (Preece, 2018, p. 67) (Singh, Sengupta, & Lakshminarayanan, 2020, p. 11).

The idea behind textual justifications is to provide the explanation of the model results of interest in terms of single phrases or complete sentences (Singh, Sengupta, & Lakshminarayanan, 2020, p. 11).

Diagnostic sentence generation for classification model based on input data and embeddings of conducted predictions is an example of explanation through textual justification (Singh, Sengupta, & Lakshminarayanan, 2020, p. 11).

Explainability techniques that aim to provide textual justifications are usually applied post-hoc (Preece, 2018, p. 67) and model-specific (Singh, Sengupta, & Lakshminarayanan, 2020, p. 11).

Transparent model types

Model types: *algorithmic transparency, interpretable models, intrinsic explainability, intrinsic interpretable Grey-Box ensemble model, design transparency, transparent models*.

Transparent model types are distinguished as a separate category of explainability techniques among others by (Rosenfeld & Richardson, 2019).

Transparent models that possess understandable logic as per their type allow explanations of their results to be directly derived without requiring any further adaptations to be conducted (Rosenfeld & Richardson, 2019, p. 9).

Some of the transparent model types are decision trees, linear models, nearest-neighbour models (Rosenfeld & Richardson, 2019, p. 9). Decision trees are often seen as some of the most understandable model types, since their hierarchical structure facilitates to easily understand the process of reaching specific results from particular input, in addition to fast identification of the most important features, of second most import features, etc. (Rosenfeld & Richardson, 2019, p. 10).

Visual techniques

Techniques: *activation maps, analysis of layers of a 3D-CNN using Gaussian mixture model (GMM) and binary encoding of training and test images based on their GMM components for returning similar 3D images, attention-based model, attention heatmaps, attention maps, attention mask weights, AttentiveChrome NN, RETAIN (REverse Time AttentIoN), class activation mapping, class maps, color based nomogram, explicitly capturing and displaying the interactions learned by a neural network, explanation maps, graphing the functional relationship between the predicted response and the feature for individual observations, histogram, HistoTrend, Interactive Training, Interactive Visualization, neural activation visualization, neural interaction detection, Neural Interpretation Diagram (NID), neural rating and tips generation, Neurons Activation (NA), Node-Link Vis, Partial Dependency Plots (PDP), visualizations, heatmaps, visual comparative analysis, visual word constraint, visualize convolutional filters, visualize filters and activations, visualize the activations of each layer of a trained CNN, visualize the decision boundary in a two-dimensional plane, visualize the discrimination of data cohorts by means of projections guided by paths through the data (tours), visualize the effect of individual inputs to the output, visualize the features of the different layers by regularized optimization in image space, visualize what computations and neuron activations occur in the intermediate layers of deep neural networks, visible NNs, auditing, CLEAR (CLass-Enhanced Attentive Response), Variable Interaction Network (VIN), Dead Weight (DeadWeight), MinMax (DeadWeight), Saturated Weight (SaturatedWeight), EG (Expressive gradients), Forest Floor, Prospector, Orthogonal Projection of Input Attributes (OPIA), U-Net based architecture and key points, U-Net with shape attention stream, mapping between image to reports, image reconstruction, Info Flow, Information Plane, mask perturbation, SAUNet, self-organizing maps,*

*show the dataflow through the computational graph, Show-and-Tell Explanations, SmoothGrad, SmoothGrad saliency maps, susceptibility maps, grad*input, gradient weighted class activation mapping (GradCAM), gradient-based, GradHM+AS, GradHM+TS, gradient-weighted heatmap (GradHM), Guided Grad-CAM, Guided-GradHM, Guided-GradHM+AS, Guided-GradHM+TS, integrated gradients*.

Visual techniques are distinguished as a separate category of explainability techniques among others by (Preece, 2018, p. 67).

Visualization techniques aim to establish the connection in a graphical form between, usually, a subset of features and the model to be explained (Rosenfeld & Richardson, 2019, p. 12). Hence, the focus is not on exact understanding of the model's logic, but rather to just visually justify and to persuade about the correctness of the results (Rosenfeld & Richardson, 2019, p. 13).
A well-known visual explanations technique is to provide the so-called Partial Dependency Plots (PDP) that visualize the outcome against a specific subset of the input, hence supporting to better understand the dependency between that specific subset of the input and the outcome (Guidotti, et al., 2019, p. 17).

Explainability techniques that primarily aim to provide a specific visualization as an explanation are usually applied post-hoc (Preece, 2018, p. 67), model-specific or model-agnostic (Carvalho, Pereira, & Cardoso, 2019), and able to explain globally the entire model or a specific local outcome (Rosenfeld & Richardson, 2019, p. 5).

6.3 Taxonomy relational structure

[A11] Identify and describe the relationships. In order to establish relations between categories of neighbouring levels (i.e., credit scoring model components to approaches of social media user profiling in credit scoring on the one hand, and approaches for social media user profiling in credit scoring to explainability approaches for social media user profiling in credit scoring on the other hand), the respective information from the selected publications from three conducted SLRs in chapter 5 is extracted. The complete taxonomy view is depicted on Figure 17.

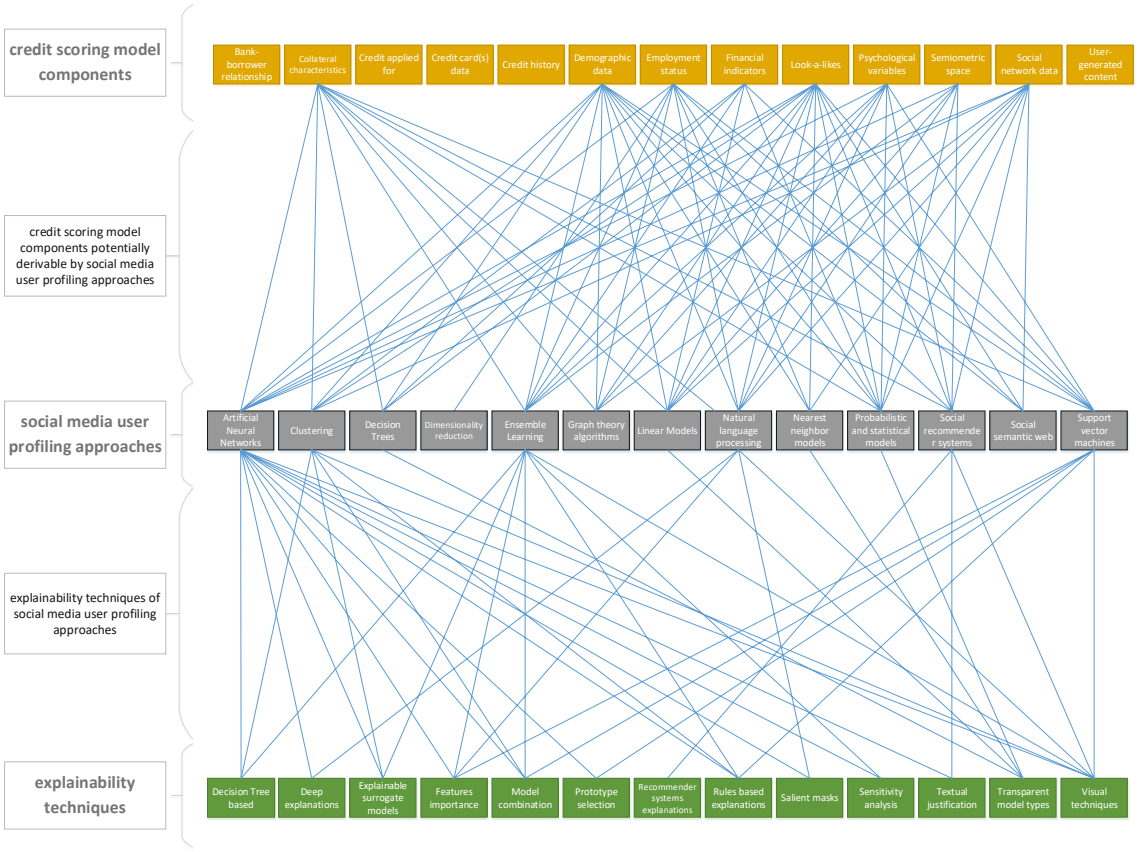

Figure 17. Taxonomy complete view

6.3.1 Credit scoring model components to social media user profiling

Under the consideration of which social media user profiling approaches for deriving which credit scoring model components are applicable (as of selected publication from the conducted SLRs), the respective relations between credit scoring model components and approaches for social media user profiling are established. In other words, in order for a relation between a specific category of credit scoring model components and a specific category of social media user profiling approaches to exist the following conditions should be met: membership of a particular credit scoring model component in a specific category of credit scoring model components, membership of a particular social media user profiling approach in a specific category of social media user profiling approaches, existence of evidence that this particular social media user profiling approach is applicable to derive this particular credit scoring component.

The complete referenced overview of the relations between credit scoring model components to approaches for social media user profiling is provided in Appendix M. Credit scoring model

components to social media user profiling. The following subsections contain exemplified overviews for each of the established relations.

Bank-borrower relationship by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for bank-borrower relationship is shown on Figure 18.

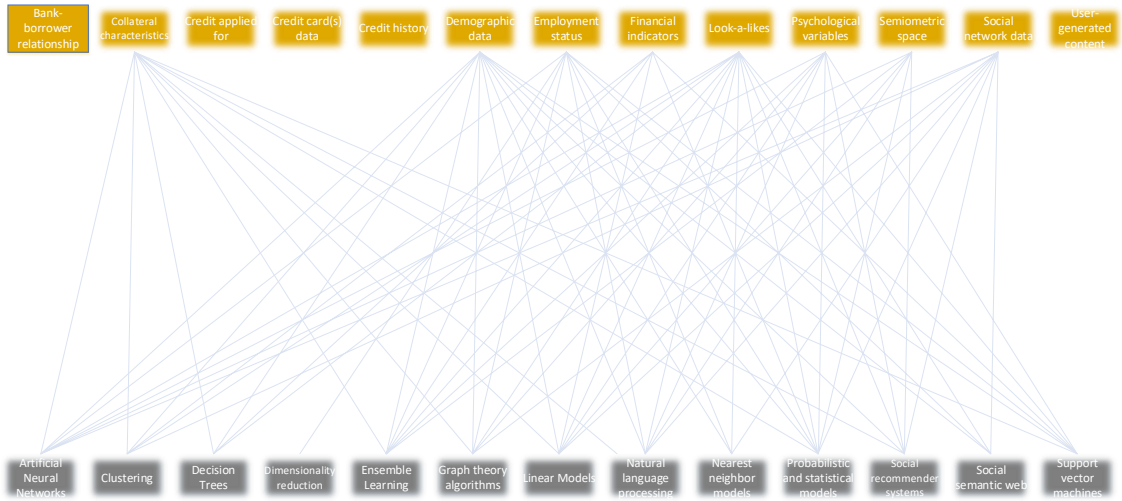

Figure 18. Taxonomy part re bank-borrower relationship using social media user profiling approaches

→ none

Collateral characteristics by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for collateral characteristics is shown on Figure 19.

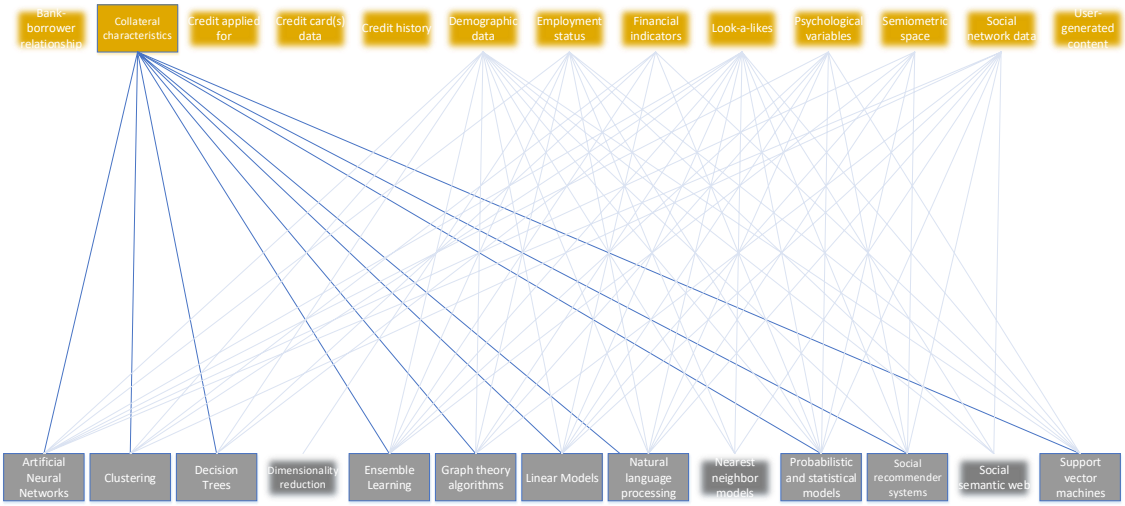

Figure 19. Taxonomy part re collateral characteristics using social media user profiling approaches

→ **Artificial neural networks**, e.g., a complex neural network model to predict geolocation based on joined text messages, user metadata and network representations (Zhang, Fu, Jiang, Bao, & Zeng, 2018, p. 5).

→ **Clustering models**, e.g., segment users into visitors and citizens for concrete locations based on their behavioural profiles obtained by means of clustering (Béjar, et al., 2016).

→ **Decision Trees**, e.g., CART tree based models as models combined in feature refinement layer to gain more effective information from textual input features in the user attributes classification task, inferring location among others (Xu, Tadesse, Fei, & Lin, 2019, p. 167).

→ **Ensemble Learning**, e.g., random forest and XGBoost as models combined in feature refinement layer to gain more effective information from textual input features in the user attributes classification task, inferring location among others (Xu, Tadesse, Fei, & Lin, 2019, p. 167).

→ **Graph Theory algorithms**, e.g., predict user's location based on social graph, for construction of which all user's followers and followees are treated as friends (Xu, Cui, Zhu, & Yang, 2014, p. 81).

→ **Linear models**, e.g., logistic regression models as models combined in feature refinement layer to gain more effective information from textual input features in the user attributes classification task, inferring location among others (Xu, Tadesse, Fei, & Lin, 2019, p. 167).

→ **Natural Language Processing**, e.g., leverage language models to determine user's location from shared content (Xu, Cui, Zhu, & Yang, 2014, p. 77).

→ **Probabilistic and statistical models**, e.g., residence location inference based on probability of friendship versus distance between users, probability of two users sharing the same residence location versus their social proximity (i.e., percentage of common friends), and two users sharing the same residence location versus their content proximity (i.e., the similarity of their generated content) (Xu, Cui, Zhu, & Yang, 2014).

→ **Social Recommender Systems**, e.g., location-based recommendations that facilitate location identification through utilizing points of interest (POIs) in a hierarchical tree-based structure (Ta, Li, Hu, & Feng, 2019, p. 1).

→ **Support Vector Machines**, e.g., infer region of users by SVM utilizing user profile joint model considering shared content and social network of users (Xu, et al., 2019, pp. 1-3).

Credit applied for by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for credit applied is shown on Figure 20.

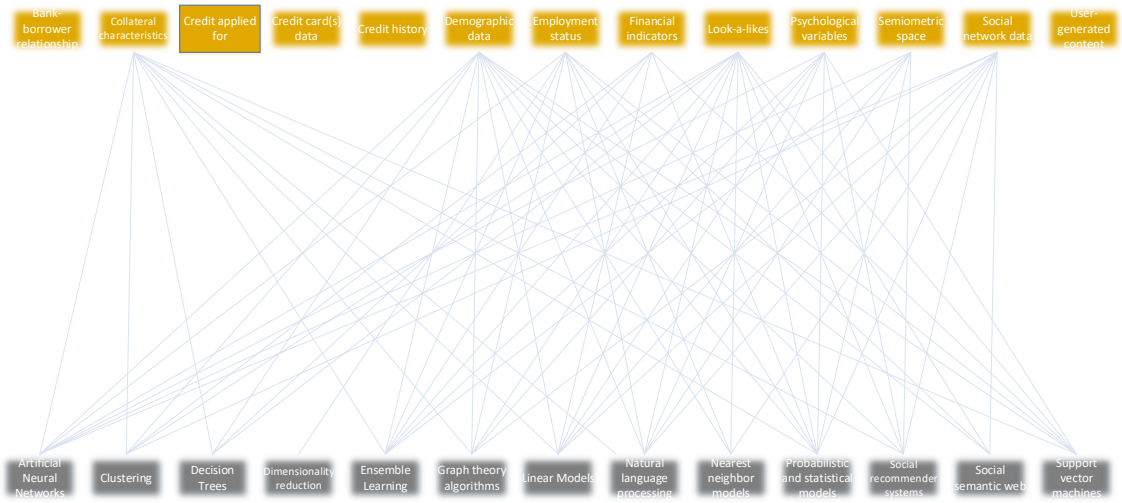

Figure 20. Taxonomy part re credit applied for using social media user profiling approaches

→ none

Credit card(s) data by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for credit card(s) data is shown on Figure 21.

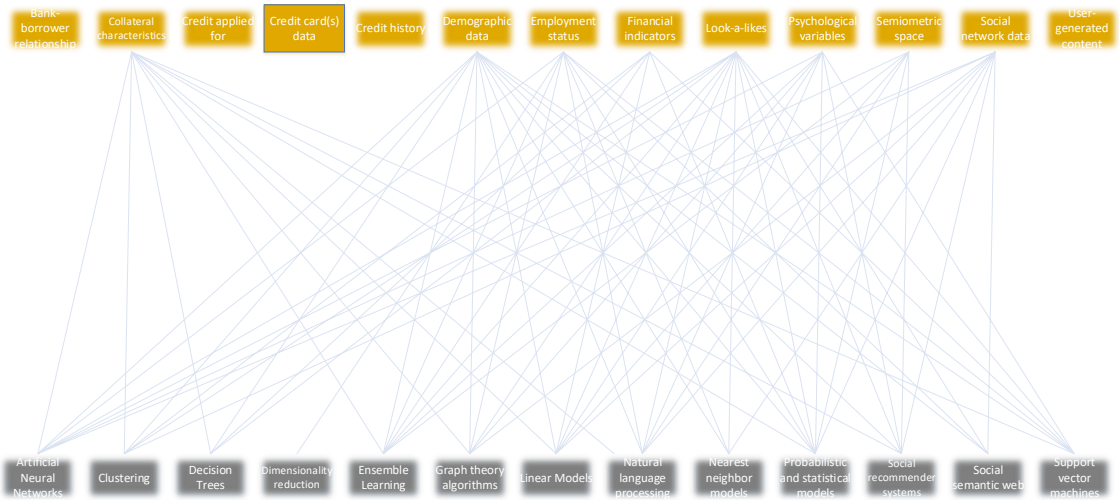

Figure 21. Taxonomy part re credit card(s) data using social media user profiling approaches

→ none

Credit history by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for credit history is shown on Figure 22.

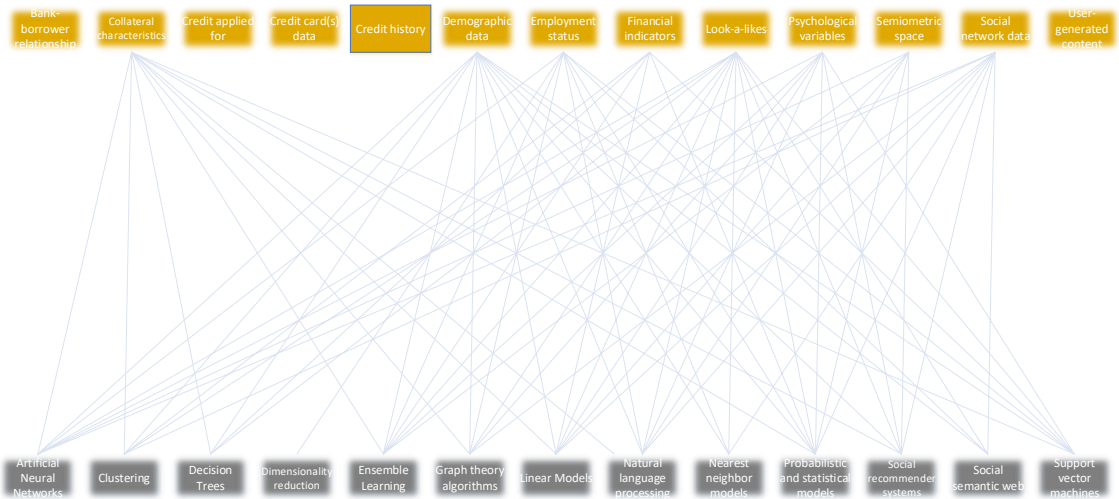

Figure 22. Taxonomy part re credit history using social media user profiling approaches

→ none

Demographic data by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for demographic data is shown on Figure 23.

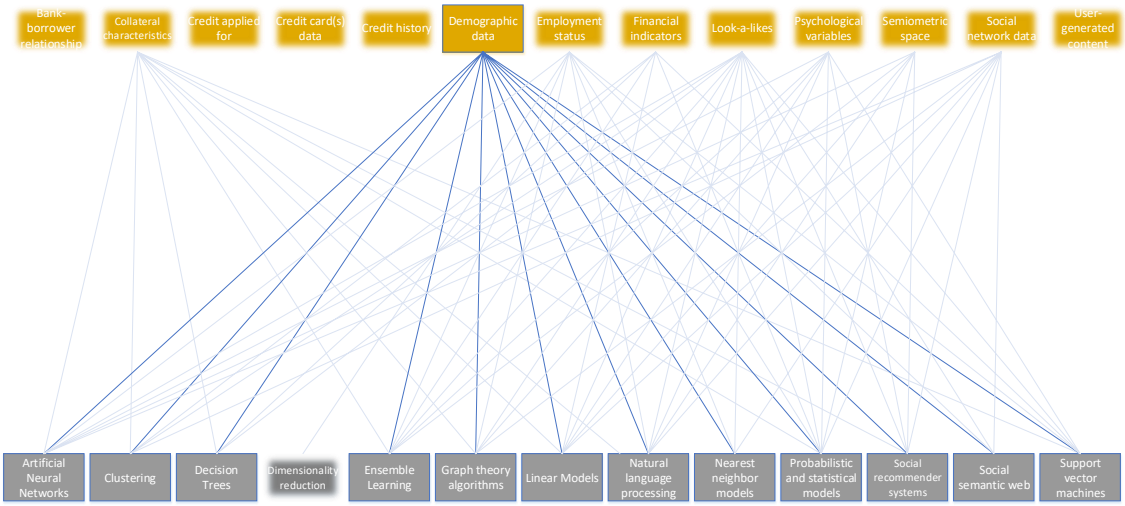

Figure 23. Taxonomy part re demographic data using social media user profiling approaches

→ **Artificial Neural Networks**, e.g., a joint learning model with LSTM (Long Short-Term Memory) model to distinguish users' ages and gender (Zhang, Fu, Jiang, Bao, & Zeng, 2018, p. 5).

→ **Clustering models**, e.g., predict age of social media users based on clustering similar user profiles (De Salve, Guidi, Ricci, & Mori, 2018).

→ **Decision trees**, e.g., determine hoaxes, frauds, and authorship, hence identity verification, of online users by decision trees (Peng, Detchon, Choo, & Ashman, 2017, p. 2).

→ **Ensemble learning**, e.g., gender prediction of social media users by various ensemble classifiers (Hirt, Kühl, & Satzger, 2019).

→ **Graph theory algorithms**, e.g., utilize various algorithms on graphs (in particular search on graphs) applied to an extended social graph as part of determining age, gender, marital status, and even weight with height of social media users (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015).

→ **Linear models**, e.g., to predict users' ages using logistic regression based on extracted stylistic features and lexical features from generated content (Zhang, Fu, Jiang, Bao, & Zeng, 2018, p. 5).

→ **Natural Language Processing**, e.g., determine age and gender from the particular language usage by social media users in their shared textual content (Li, et al., 2019, p. 274).

→ **Nearest neighbour models**, e.g., utilize neighbourhood-based information from an extended social graph as part of determining age, gender, marital status, and even weight with height of social media users (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015).

→ **Probabilistic and statistical models**, e.g., Naïve Bayes and its various modifications (such as Relational Naïve Bayes and Collective Naïve Bayes) to determine age, gender, marital status, and even weight with height of social media users (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015).

→ **Social recommender systems**, e.g., predict age of users based on movie preferences and recommendations (De Salve, Guidi, Ricci, & Mori, 2018).

→ **Social semantic web**, e.g., utilize extracted contextual semantic representation of text on social media as part of a system to construct user profile containing, among others, age and gender (Li, et al., 2019, p. 273).

→ **Support vector machines**, e.g., to predict users' ages using SVM based on extracted stylistic features and lexical features from generated content (Zhang, Fu, Jiang, Bao, & Zeng, 2018, p. 5).

Employment status by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for employment status is shown on Figure 24.

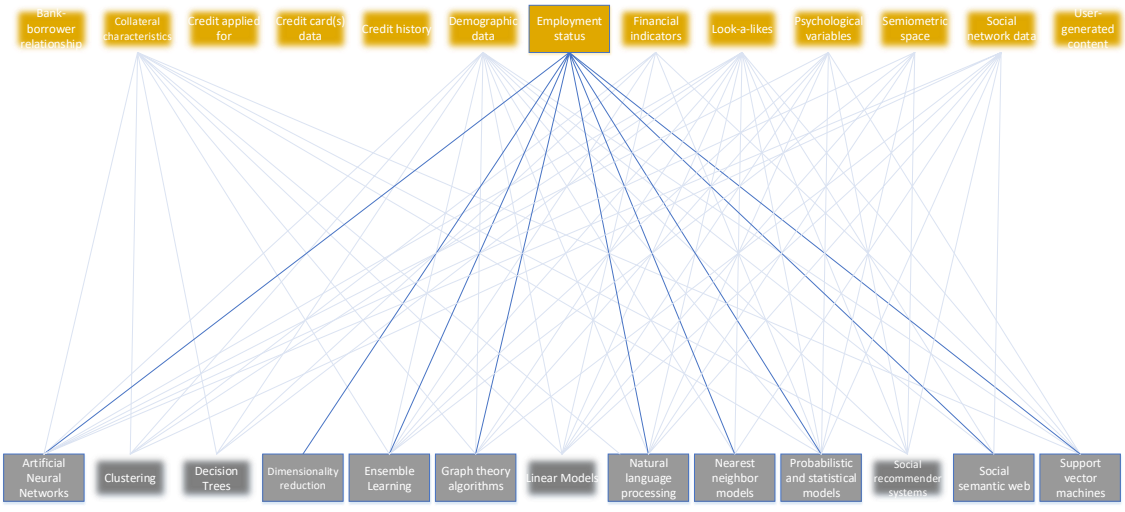

Figure 24. Taxonomy part re employment status using social media user profiling approaches

→ **Artificial Neural Networks**, e.g., a joint learning model with LSTM (Long Short-Term Memory) model to distinguish users' professions (Zhang, Fu, Jiang, Bao, & Zeng, 2018, p. 5).

→ **Dimensionality reduction**, e.g., reduce social network feature into a low dimensional space as a preliminary step for solving occupation prediction problem (Tong, Yao, Wang, & Yang, 2016).

→ **Ensemble learning**, e.g., apply stacking of the results of various machine learning approaches to address occupation prediction task (Tong, Yao, Wang, & Yang, 2016).

→ **Graph theory algorithms**, e.g., apply graph embedding approach of user's social network as a preliminary step for solving occupation prediction problem (Tong, Yao, Wang, & Yang, 2016).

→ **Natural Language Processing**, e.g., transform text information in shared content to latent representation using, among others, term frequency and inverse document frequency criterion, and utilize language model word2vec as an intermediate step for solving occupation prediction problem (Tong, Yao, Wang, & Yang, 2016).

→ **Nearest neighbour models**, e.g., construct user profile including, among others, occupation prediction using *k*-Nearest Neighbours approach (Li, et al., 2019, pp. 281-282).

→ **Probabilistic and statistical models**, e.g., employ a modified probabilistic model for automatically extracting representative words and identify users' latent topic distribution from their shared content as parts of method to predict social media users' occupation (Huang, Yu, Wang, & Cui, 2015).

→ **Social semantic web**, e.g., utilize extracted contextual semantic representation of text on social media as part of a system to construct user profile containing, among others, occupation (Li, et al., 2019, p. 273).

→ **Support vector machines**, e.g., construct user profile including, among others, occupation prediction using SVM (Li, et al., 2019, pp. 281-282).

Financial indicators by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for financial indicators is shown on Figure 25.

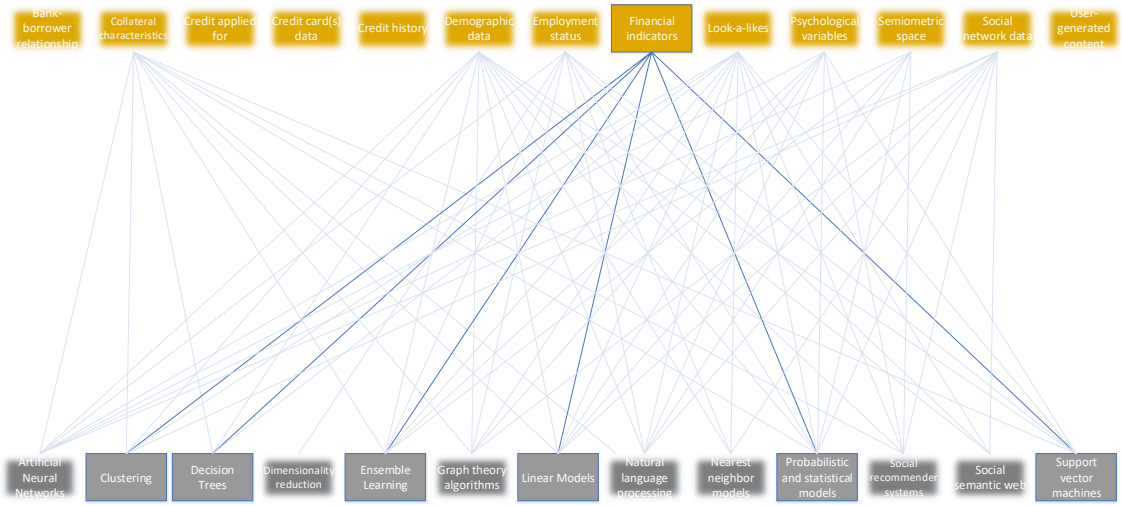

Figure 25. Taxonomy part re financial indicators using social media user profiling approaches

→ **Clustering models**, e.g., apply spectral clustering as part of method to infer socio-economic status of social media users (Lampos, Aletras, Geyti, Zou, & Cox, 2016, p. 691).

→ **Decision trees**, e.g., construct Decision Tree as one of the Tier-1 classifiers to provide extended features in the two-tier system to learn credit labels from social media data (Guo, et al., 2016).

→ **Ensemble learning**, e.g., apply stacking on Tier-1 classifiers and boosting on Tier-2 in the two-tier system to learn credit labels from social media data (Guo, et al., 2016).

→ **Linear models**, e.g., construct logistic regression as one of the Tier-1 classifiers to provide extended features in the two-tier system to learn credit labels from social media data (Guo, et al., 2016).

→ **Probabilistic and statistical models**, e.g., apply Naïve Bayes as one of the Tier-1 classifiers to provide extended features in the two-tier system to learn credit labels from social media data (Guo, et al., 2016).

→ **Support vector machines**, e.g., develop SVM as one of the Tier-1 classifiers to provide extended features in the two-tier system to learn credit labels from social media data (Guo, et al., 2016).

Look-a-likes by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for look-a-likes is shown on Figure 26.

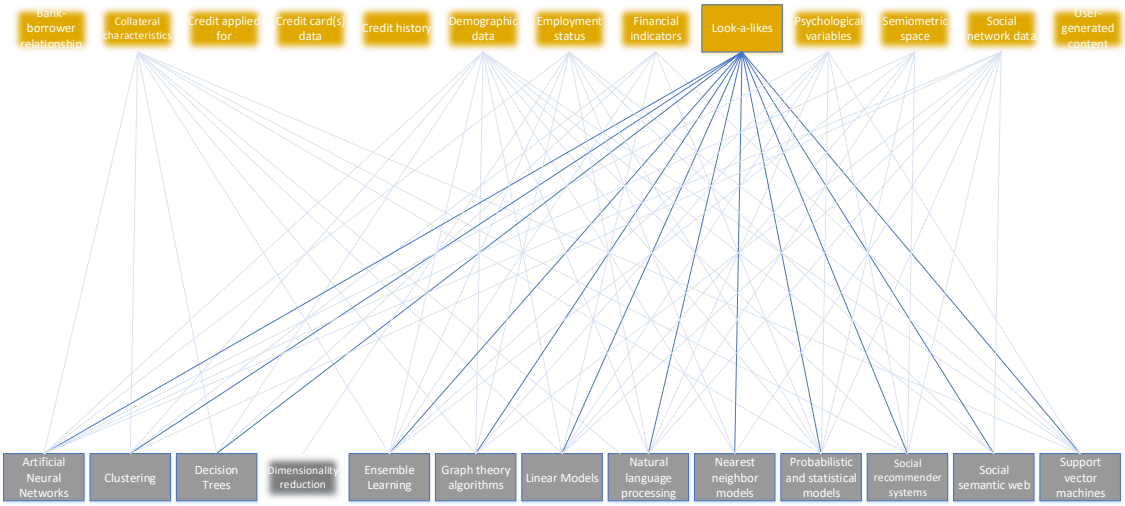

Figure 26. Taxonomy part re look-a-likes using social media user profiling approaches

→ **Artificial Neural Networks**, e.g., apply various deep neural networks (e.g., Convolutional Neural Network, bi-directional Long-Short Term Memory) as classifiers to classify similar users belonging to specific categories in the public health domain (Zhang & Bors, 2019).

→ **Clustering models**, e.g., cluster social media users with respect to their age or similar movie preferences using k-means (De Salve, Guidi, Ricci, & Mori, 2018).

→ **Decision trees**, e.g., use decision trees to classify similar user on their different labels, e.g., political orientation (Chen, Zhu, Guo, & Liu, 2014, p. 163).

→ **Ensemble learning**, e.g., apply random forest as one of the classifiers to classify similar users belonging to specific categories in the public health domain (Zhang & Bors, 2019).

→ **Graph theory algorithms**, e.g., apply graph-based community discovery method, i.e., determine communities of similar users whom same particular labels are to be assigned (Chen, Zhu, Guo, & Liu, 2014, p. 163).

→ **Linear models**, e.g., apply logistic as one of the classifiers to classify similar users belonging to specific categories in the public health domain (Zhang & Bors, 2019).

→ **Natural Language Processing**, e.g., employ a language model based on sentiment classification approach to compute the relationship strength between users, accounting for similarity between users (Ju & Tao, 2017).

→ **Nearest neighbor models**, e.g., social voting within users' neighborhoods as component to find user's potential interests, hence facilitating look-a-likes identification (Eke, Norman, Shuib, & Nweke, 2019, p. 144917).

→ **Probabilistic and statistical models**, e.g., apply probabilistic model in a heterogeneous network to identify similar users whom to propagate specific interests implied by similarity (Chen, Zhu, Guo, & Liu, 2014, p. 163).

→ **Social recommender systems**, e.g., content-based or collaborative filtering to match user's interests respectively depending on interest in similar content items or similar groups of users by interests, hence facilitating look-a-likes identification (Eke, Norman, Shuib, & Nweke, 2019, p. 144917).

→ **Social semantic web**, e.g., an ontology-based user profile acquisition (OUPA) method to automatically construct and maintain user ontology of personal interests, hence facilitating look-a-likes identification (Eke, Norman, Shuib, & Nweke, 2019, p. 144916).

→ **Support vector machines**, e.g., use SVM to classify similar users by their gender, age, regional origin and political orientation (Chen, Zhu, Guo, & Liu, 2014, p. 163).

Psychological variables by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for psychological variables is shown on Figure 27.

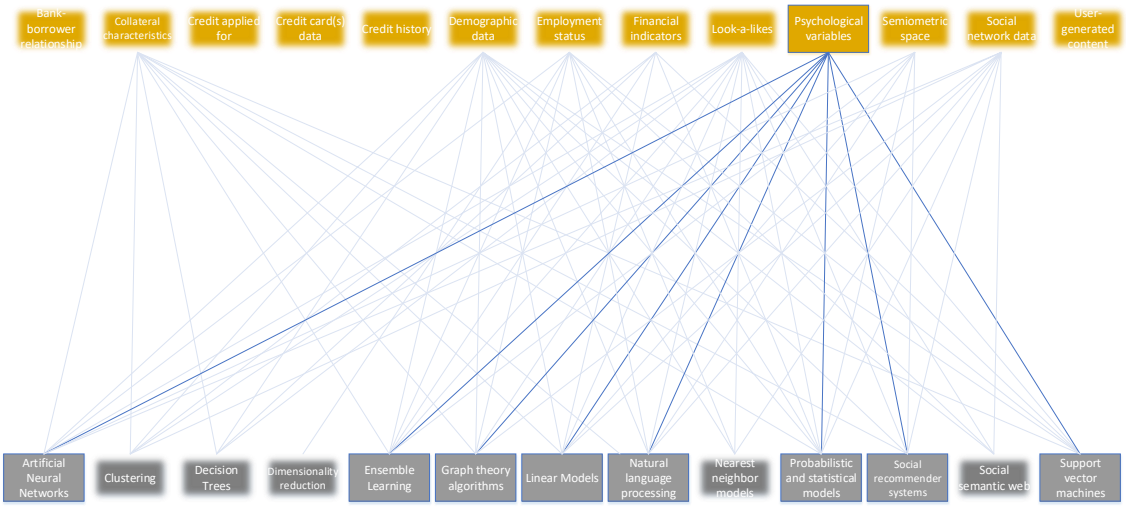

Figure 27. Taxonomy part re psychological variables using social media user profiling approaches

→ **Artificial Neural Networks**, e.g., apply Long Short Term Memory (LSTM) neural networks on multisource multi-modal temporal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to various psychological variables identification (Baklouti, 2014a).

→ **Ensemble learning**, e.g., apply gradient boosting on multisource multi-modal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to various psychological variables identification (Baklouti, 2014a).

→ **Graph theory algorithms**, e.g., apply Graph Theoretic Analysis as part of approach on social media data to predict specific psychosocial traits (Kandias, Mitrou, Stavrou, & Gritzalis, 2014).

→ **Linear models**, e.g., apply logistic regression on multisource multi-modal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to various psychological variables identification (Baklouti, 2014a).

→ **Natural Language Processing**, e.g., extract and incorporate various linguistic and LDA features as part of approach on multisource multi-modal temporal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to various psychological variables identification (Baklouti, 2014a).

→ **Probabilistic and statistical models**, e.g., apply Naïve Bayes on multisource multi-modal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to various psychological variables identification (Baklouti, 2014a).

→ **Social recommender systems**, e.g., make use of non-negative matrix factorization (NMF), largely used in recommender systems, as part of approach on multisource multi-modal temporal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to various psychological variables identification (Baklouti, 2014a).

→ **Support vector machines**, e.g., apply SVM for comment classification as part of approach on social media data to predict specific psychosocial traits (Kandias, Mitrou, Stavrou, & Gritzalis, 2014).

Semiometric space by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for semiometric space is shown on Figure 28.

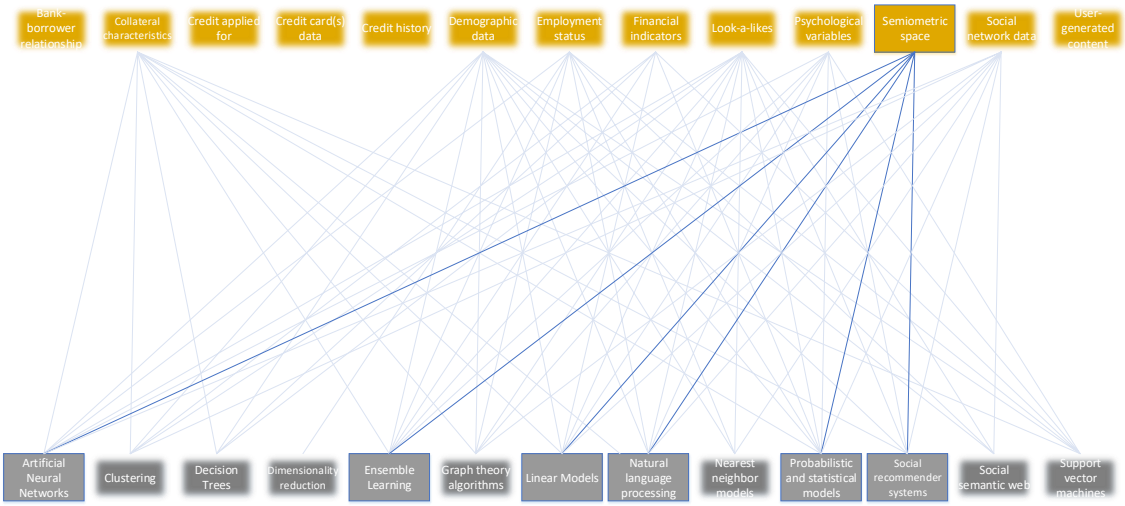

Figure 28. Taxonomy part re semiometric space using social media user profiling approaches

→ **Artificial Neural Networks**, e.g., apply Long Short Term Memory (LSTM) neural networks on multisource multi-modal temporal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to construct semiometric space projection (Liberati & Camillo, 2018).

→ **Ensemble learning**, e.g., apply gradient boosting on multisource multi-modal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to construct semiometric space projection (Liberati & Camillo, 2018).

→ **Linear models**, e.g., apply logistic regression on multisource multi-modal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to construct semiometric space projection (Liberati & Camillo, 2018).

→ **Natural Language Processing**, e.g., extract and incorporate various linguistic and LDA features as part of approach on multisource multi-modal temporal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit

psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to construct semiometric space projection (Liberati & Camillo, 2018).

→ **Probabilistic and statistical models**, e.g., apply Naïve Bayes on multisource multi-modal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to construct semiometric space projection (Liberati & Camillo, 2018).

→ **Social recommender systems**, e.g., make use of non-negative matrix factorization (NMF), largely used in recommender systems, as part of approach on multisource multi-modal temporal data from social media to conduct user personality profiling according to the MBTI typology, i.e., exhibit psychological preferences on how people make decisions and perceive the world (Buraya, Farseev, & Filchenkov, 2018), respectively appliable to construct semiometric space projection (Liberati & Camillo, 2018).

Social network data by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for social network data is shown on Figure 29.

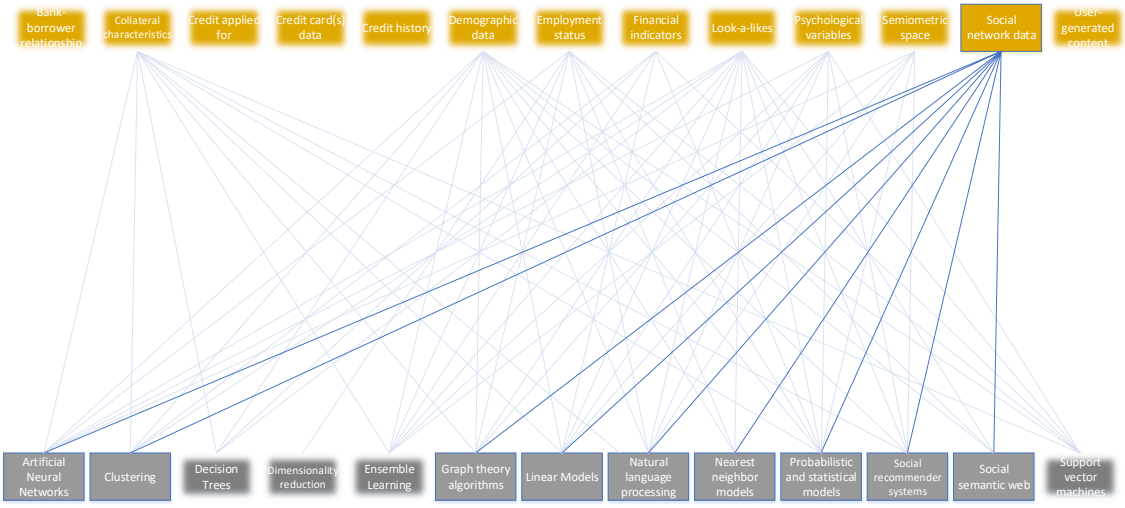

Figure 29. Taxonomy part re social network data using social media user profiling approaches

→ **Artificial Neural Networks**, e.g., use deep learning techniques of Network Representation Learning to identify social connections among the users for their social network construction (C C & Mohan, 2019, p. 1938).

→ **Clustering models**, e.g., apply various clustering techniques to determine user's realistic neighbours' network from user's complete social network (C C & Mohan, 2019, p. 1940).

→ **Graph theory algorithms**, e.g., apply network embedding approaches to identify social connections among the users for their social network construction (C C & Mohan, 2019, p. 1938).

→ **Linear models**, e.g., to use logistic regression to learn the parameters of the Markov random field, which in turn is used to model the relations in an interaction network (Chen, Zhu, Guo, & Liu, 2014, p. 164).

→ **Natural Language Processing**, e.g., use language modelling techniques, inspired by generalization of NLP, to learn specific relationships between nodes to construct the respective network (C C & Mohan, 2019, p. 1940).

→ **Nearest neighbour models**, e.g., determine user's trustable social neighbours' network from user's complete social network (C C & Mohan, 2019).

→ **Probabilistic and statistical models**, e.g., apply random walk based embedding techniques, such as hierarchical representation learning for networks (HARP), for network embedding and to identify social relations (C C & Mohan, 2019, p. 1948).

→ **Social recommender systems**, e.g., utilize autoencoder based social recommender system (AESR) to extracts the network structure of user-user interactions from user-item interactions (C C & Mohan, 2019, p. 1940).

→ **Social semantic web**, e.g., incorporate notion of semantic friends for the social user-user interaction network construction based on identified top-k semantic friends of each user (C C & Mohan, 2019, p. 1938).

User-generated content by social media user profiling approaches

The part of the taxonomy depicting explainability techniques for user-generated content is shown on Figure 30.

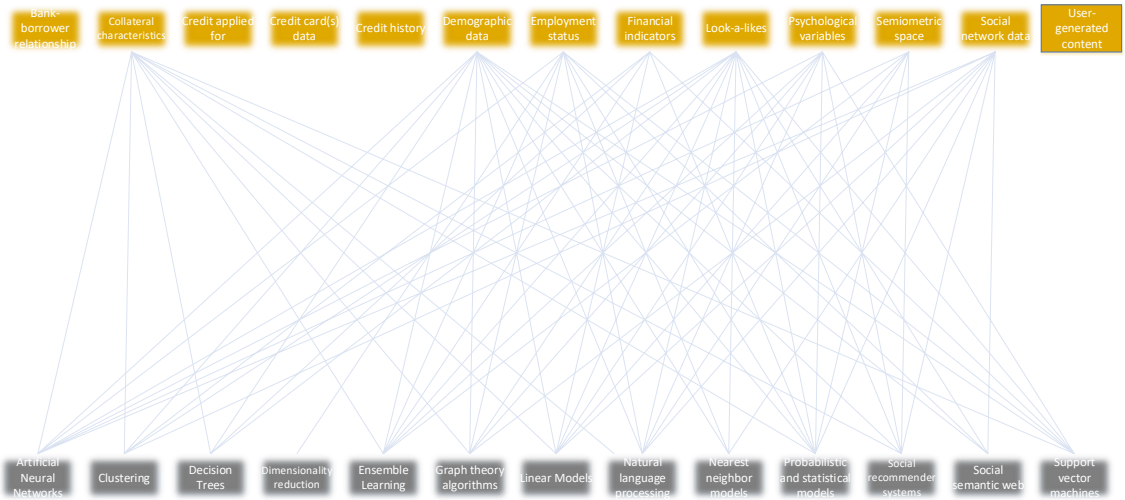

Figure 30. Taxonomy part re user-generated content using social media user profiling approaches

→ none

6.3.2 Social media user profiling approaches to explainability techniques

Under the consideration of which explainability techniques are applicable to which social media user profiling approaches (as of selected publication from the conducted SLRs), the respective relations between social media user profiling approaches and explainability techniques thereof are established. In other words, in order for a relation between a specific category of social media user profiling approaches and a specific category of explainability techniques to exist the following conditions should be met: membership of a particular social media user profiling approach in a specific category of social media user profiling approaches, membership of a particular explainability technique in a specific category of explainability techniques, existence of evidence that this particular explainability technique is applicable to explain this particular social media user profiling approach.

The complete referenced overview of the relations between approaches for social media user profiling to explainability of approaches for social media user profiling is provided in Appendix N. Social media user profiling approaches to explainability techniques. The following subsections contain exemplified overviews for each of the established relations.

Explainability techniques for Artificial Neural Networks

The part of the taxonomy depicting explainability techniques for Artificial Neural Networks is shown on Figure 31.

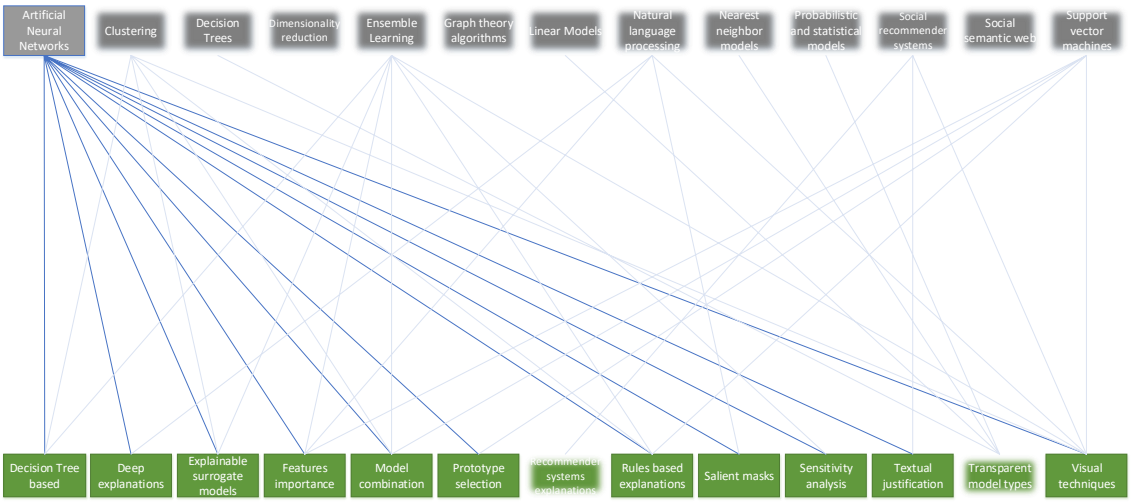

Figure 31. Taxonomy part re explainability techniques for ANN

→ **Decision Tree based explanations**, e.g., approximating the ANN model with Single Tree Approximation (STA) (Guidotti, et al., 2019, pp. 22-23).

→ **Deep explanations**, e.g., invoking back propagation and/or activation techniques to generate a class activation mapping, i.e., local explanations for a particular prediction outcome label (Guidotti, et al., 2019, p. 30).

→ **Explainable surrogate models**, e.g., explaining an image classification decision of the respective deep learning models by the Local Interpretable Model-agnostic Explanations (LIME) (Preece, 2018, p. 66).

→ **Features importance**, e.g., provide explainability of neural networks based on set of Quantitative Input Influence (QII) measures that capture how much input features impact output (Guidotti, et al., 2019, p. 31).

→ **Model combination**, e.g., combining image classification ANN model with explanations generation model in a coherent dual neural network system (Preece, 2018, p. 67).

→ **Prototype selection**, e.g., selecting prototypes as part of a multi-step method consisting of initial prototypes generation using genetic programming, constraining initial prototypes using input features dataset, and selecting the best prototypes for further processing to generate explanations for trained in advance neural networks (Guidotti, et al., 2019, p. 23).

→ **Rules based explanations**, e.g., explain the behaviour of ANN by the set of conjunctive, i.e., m-of-n, rules through a technique of transforming search problem of rule extraction into a learning problem (Guidotti, et al., 2019, p. 26).

→ **Salient masks**, e.g., for an image caption prediction by convolutional NN with recursive NN containing LSTM highlight areas of an image responsible for each word in generated caption (Guidotti, et al., 2019, p. 29).

→ **Sensitivity analysis**, e.g., explain ANN through sensitivity analysis and Neural Interpretation Diagram (NID), assessing the importance of axon connections and input variables' contribution (Guidotti, et al., 2019, p. 32).

→ **Textual justification**, e.g., provide textual explanations by aligning deep neural network model's internal structures with specific semantic concepts representing elements of interest, thus generating natural language explanations based on neural activations within network (Gunning & Aha, 2019, p. 54).

→ **Visual techniques**, e.g., visualize label-specific weights (or gradients) and linear combination of a late layer's activations as applied to neural networks using Grad-CAM (Guidotti, et al., 2019, p. 30).

Explainability techniques for clustering

The part of the taxonomy depicting explainability techniques for clustering is shown on Figure 32.

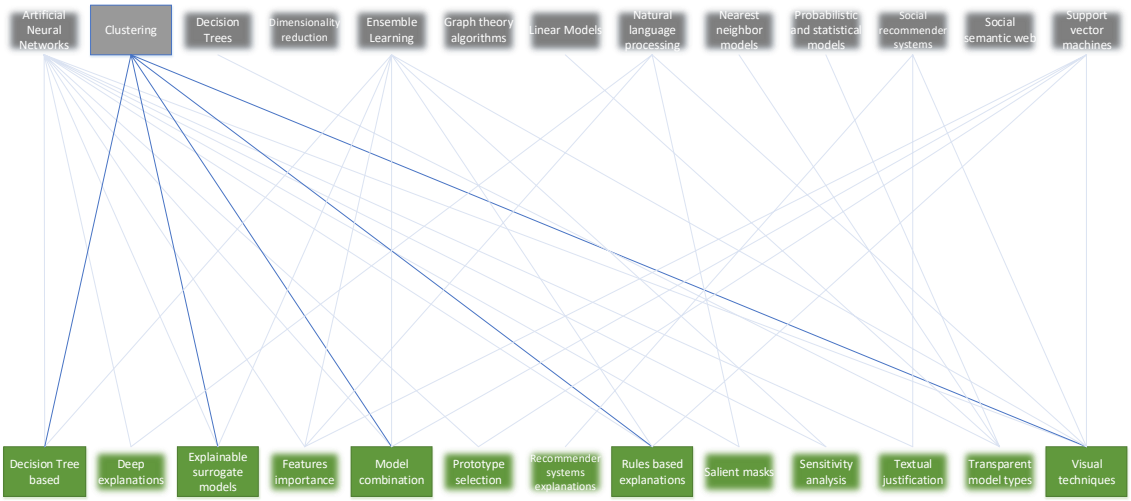

Figure 32. Taxonomy part re explainability techniques for clustering

→ **Decision Tree based explanations**, e.g., apply decision tree algorithm on data set that contains combination of input features and identified clusters (De Koninck, De Weerdt, & vanden Broucke, 2017, p. 780).

→ **Explainable surrogate models**, e.g., construct an explainable classification model to the constructed clustering model, i.e., perform reverse engineering (De Koninck, De Weerdt, & vanden Broucke, 2017, p. 780).

→ **Model combination**, e.g., apply automated similarity analysis technique through additionally generating specific similarity metrics for explanation (De Koninck, De Weerdt, & vanden Broucke, 2017, p. 779).

→ **Rules based explanations**, e.g., apply rule learning algorithm on data set that contain combination of input features and identified clusters (De Koninck, De Weerdt, & vanden Broucke, 2017, p. 780).

→ **Visual techniques**, e.g., conduct visual comparative analysis of the discovered instances respectively belonging to the different identified clusters (De Koninck, De Weerdt, & vanden Broucke, 2017, p. 779).

Explainability techniques for decision trees

The part of the taxonomy depicting explainability techniques for decision trees is shown on Figure 33.

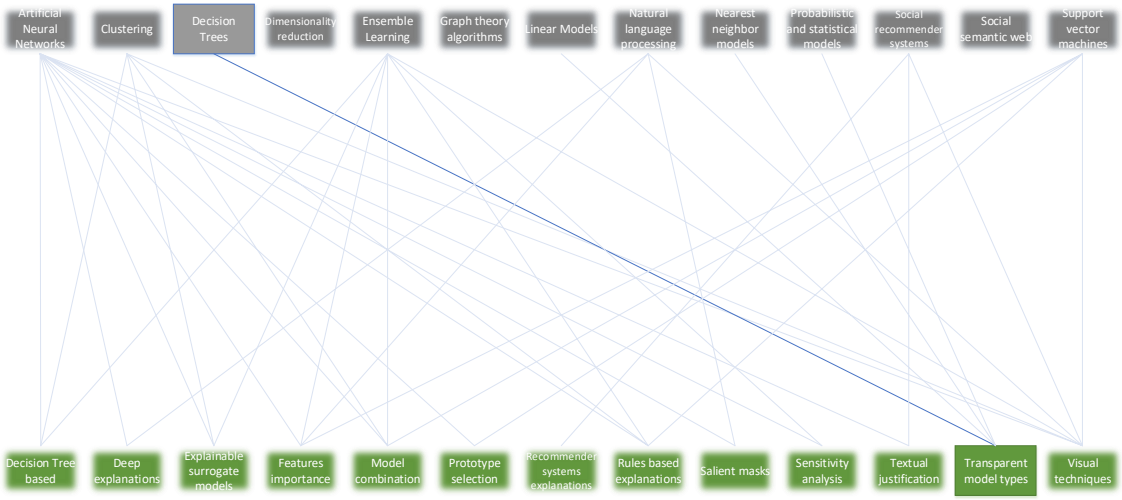

Figure 33. Taxonomy part re explainability techniques for decision trees

→ Transparent model types

Explainability techniques for dimensionality reduction

The part of the taxonomy depicting explainability techniques for dimensionality reduction is shown on Figure 34.

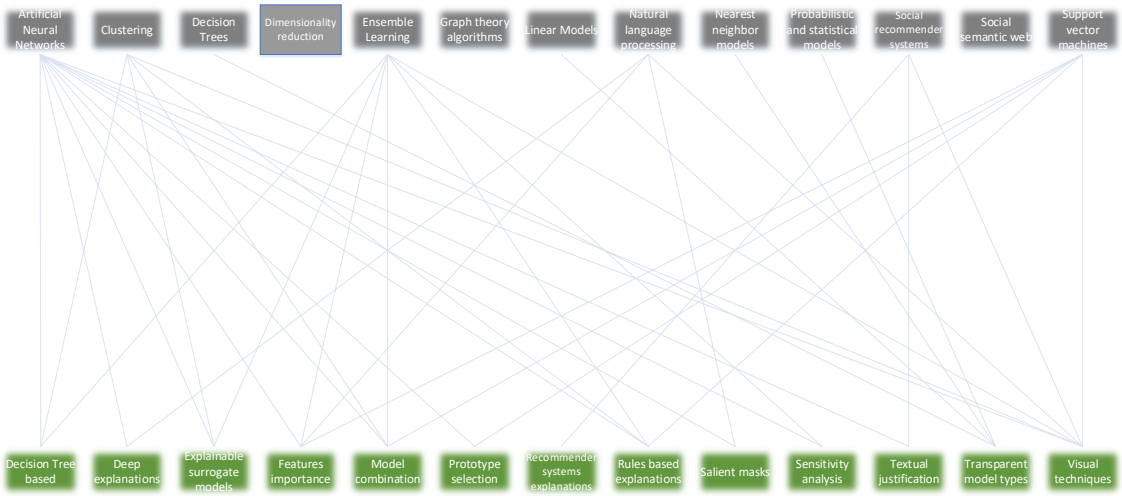

Figure 34. Taxonomy part re explainability techniques for dimensionality reduction

→ none

Explainability techniques for ensemble learning

The part of the taxonomy depicting explainability techniques for Artificial Neural Networks is shown on Figure 35.

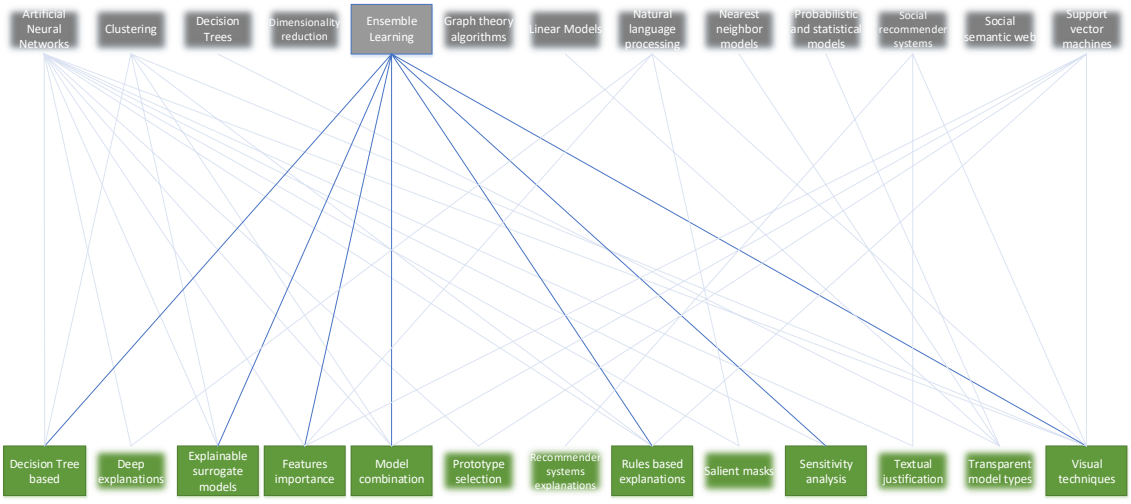

Figure 35. Taxonomy part re explainability techniques for ensemble learning

→ **Decision Tree based explanation**, e.g., approximating ensemble learning models by Single Tree Approximation (STA) (Guidotti, et al., 2019, pp. 24-25).

→ **Explainable surrogate models**, e.g., mimic a complex tree-based ATM (Additive Tree Model) using a simpler and easily explainable probabilistic generative model (Guidotti, et al., 2019, p. 25).

→ **Features importance**, e.g., estimate the contribution of input features, either positively or negatively, to target variable using SHAP values as applied to random forest, XGBoost, or other gradient boosting approaches (Ariza, Arroyo, Caparrini, & Segovia, 2020).

→ **Model combination**, e.g., combine random forest with similarity analysis between its single trees based on specific measures of dissimilarity for trees used to summarize that forest of trees through clustering, subsequently selecting archetypes of associated clusters as explanations (Guidotti, et al., 2019, p. 24).

→ **Rules based explanations**, e.g., extract the simplest and most supported decision rules form tree ensemble (like random forest) through Simplified Tree Ensemble Learner (STEL) technique (Guidotti, et al., 2019, p. 27).

→ **Sensitivity analysis**, e.g., conduct sensitivity analysis of model's input for ensemble trees addressing classification or regression tasks (Rosenfeld & Richardson, 2019, p. 13).

→ **Visual techniques**, e.g., apply visualizing technique called Forest Floor to provide an explanation for Random Forest models in form reduced higher dimensional maps of single trees to lower dimensional slices or projections with specific color codes (Käde & Von Maltzan, 2019, p. 7).

Explainability techniques for graph theory algorithms

The part of the taxonomy depicting explainability techniques for graph theory algorithms is shown on Figure 36.

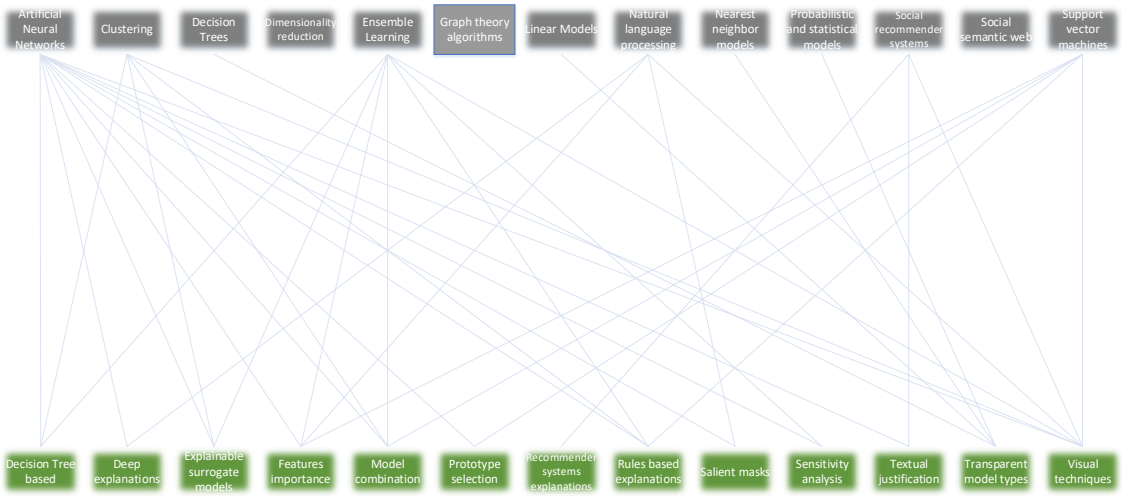

Figure 36. Taxonomy part re explainability techniques for graph theory algorithms

→ none

Explainability techniques for linear models

The part of the taxonomy depicting explainability techniques for linear models is shown on Figure 37.

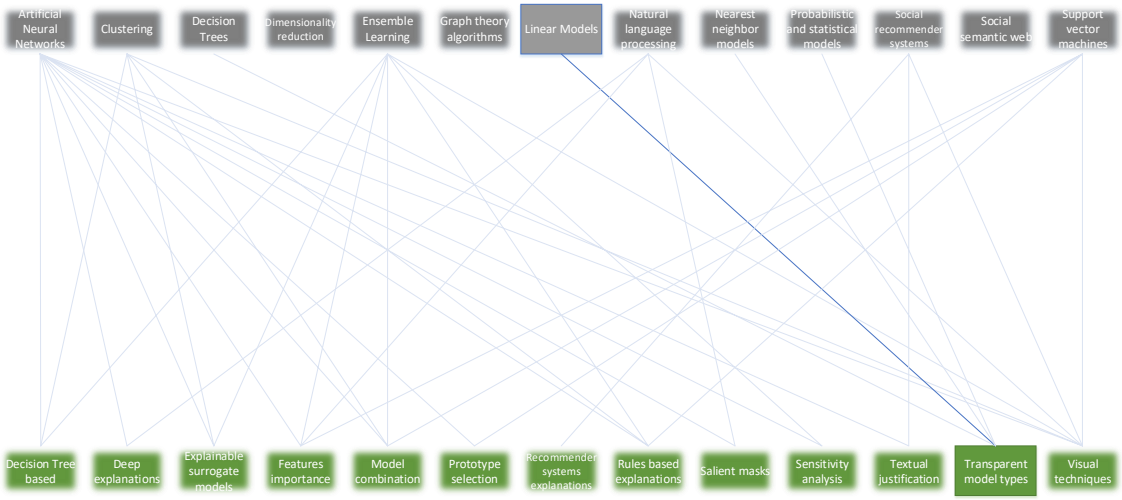

Figure 37. Taxonomy part re explainability techniques for linear models

→ Transparent model types

Explainability techniques for Natural Language Processing

The part of the taxonomy depicting explainability techniques for Natural Language Processing is shown on Figure 38.

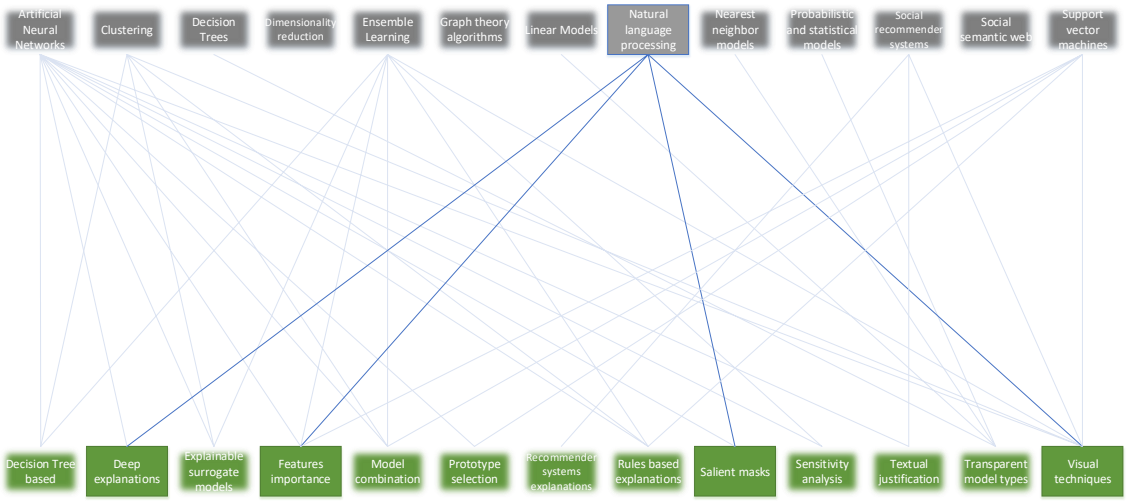

Figure 38. Taxonomy part re explainability techniques for NLP

→ **Deep explanations**, e.g., integrate backward-propagation technique from output through the model to input in order to generate explanations of an NLP model (Arras, Horn, Montavon, Müller, & Samek, 2017).

→ **Features importance**, e.g., compute scores based on layer-wise relevance propagation (LRP) that indicate how much individual features contribute to the decision of a specific language model (Arras, Horn, Montavon, Müller, & Samek, 2017).

→ **Salient masks**, e.g., identify specific words that determine particular decisions of a specific word-based model based on adaptation of the layer-wise relevance propagation (LRP) to decompose the predictions onto words (Arras, Horn, Montavon, Müller, & Samek, 2017).

→ **Visual techniques**, e.g., apply document heatmap visualizations of word-level relevance to the NLP model addressing topic categorization task (Arras, Horn, Montavon, Müller, & Samek, 2017).

Explainability techniques for nearest neighbour models

The part of the taxonomy depicting explainability techniques for nearest neighbour models is shown on Figure 39.

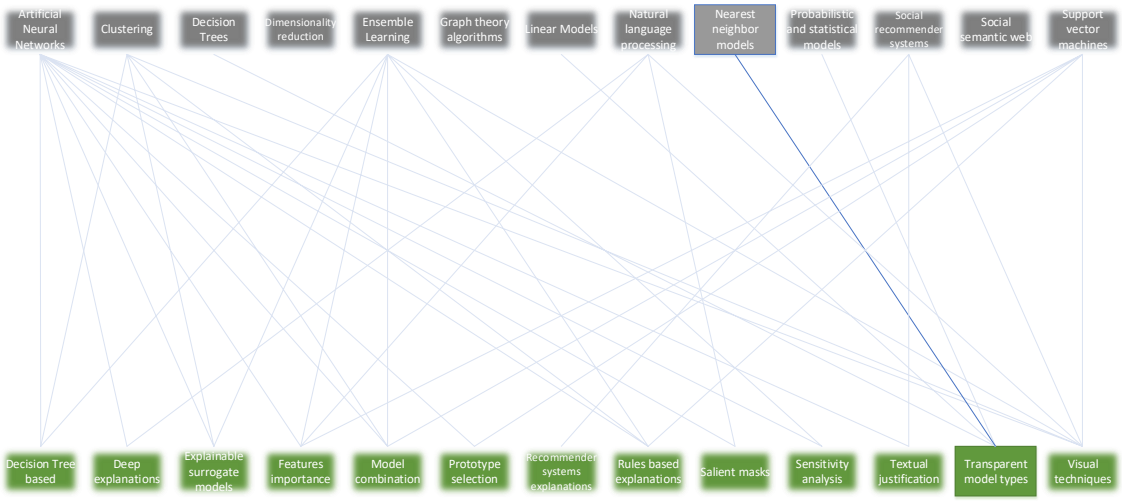

Figure 39. Taxonomy part re explainability techniques for nearest neighbour models

→ Transparent model types

Explainability techniques for probabilistic and statistical models

The part of the taxonomy depicting explainability techniques for statistical models is shown on Figure 40.

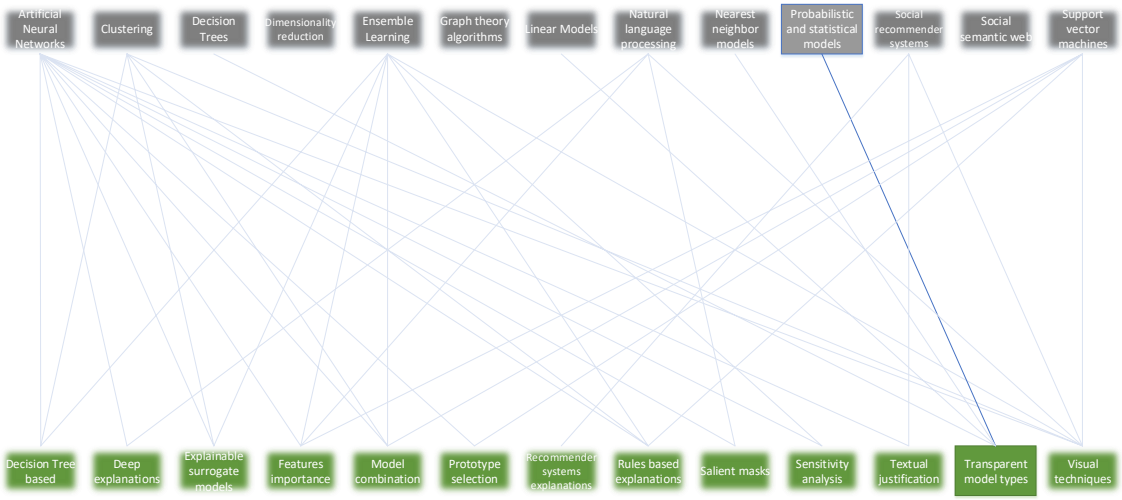

Figure 40. Taxonomy part re explainability techniques for probabilistic and statistical models

→ Transparent model types

Explainability techniques for social recommender systems

The part of the taxonomy depicting explainability techniques for social recommender systems is shown on Figure 41.

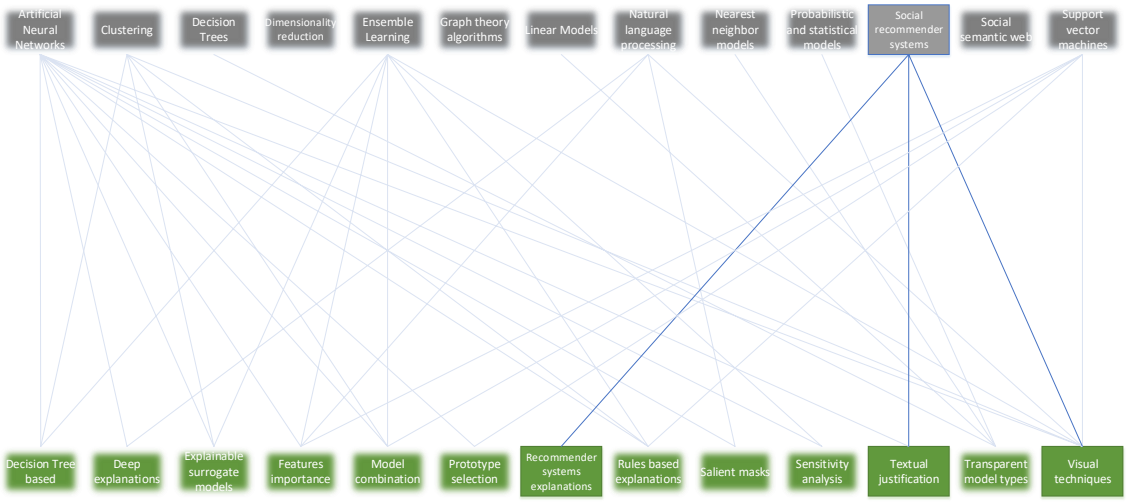

Figure 41. Taxonomy part re explainability techniques for SRS

→ **Recommender systems explanations**, e.g., apply Aspect-based Matrix Factorization model (AMF) to achieve an explainable recommendation by a collaborative decomposition of the rating matrix with the auxiliary information extracted from additional aspects (Hou, Yang, Wu, & Yu, 2019).

→ **Textual justification**, e.g., provide keywords or user-tags based explanation as a justification for recommended item in content-based recommender systems (Chen, Yan, & Wang, 2019, p. 2).

→ **Visual techniques**, e.g., provide a histogram with a grouping of neighbours in different rating categories for recommended item as an explanation of that particular recommendation in collaborative filtering (CF) systems (Chen, Yan, & Wang, 2019, p. 2).

Explainability techniques for social semantic web

The part of the taxonomy depicting explainability techniques for social semantic web is shown on Figure 42.

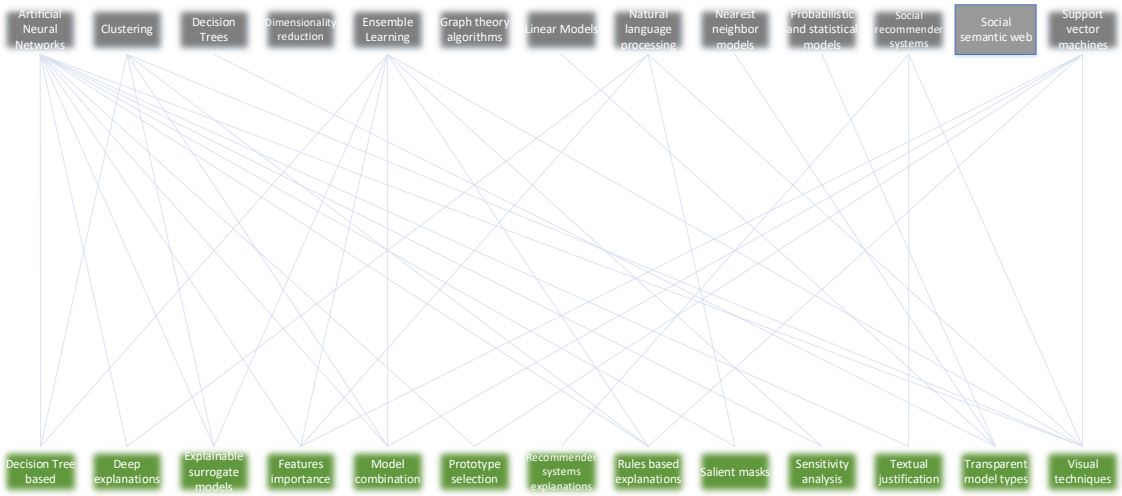

Figure 42. Taxonomy part re explainability techniques for SSW

→ none

Explainability techniques for support vector machines

The part of the taxonomy depicting explainability techniques for support vector machines is shown on Figure 43.

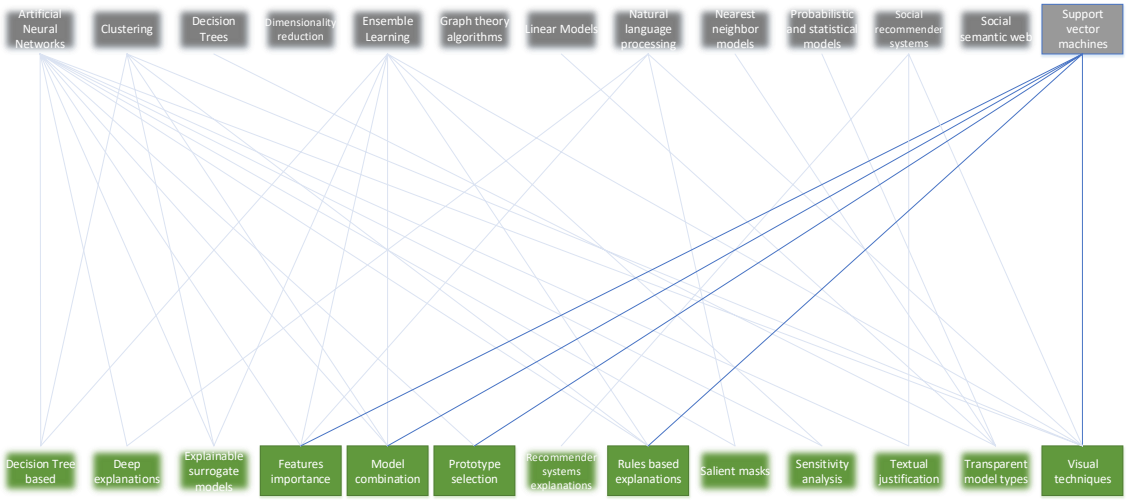

Figure 43. Taxonomy part re explainability techniques for SVM

→ **Features importance**, e.g., quantify the impact of each of the SVM model's input features as part of processing this SVM model into a colour-based nomogram representation (Belle, Calster, Huffel, Suykens, & Lisboa, 2016).

→ **Model combination**, e.g., combine the process of determining the decision function by means of SVM with clustering to identify representative instance of each class as part of the SVM+P technique proposed to generate explanations for SVM (Guidotti, et al., 2019, p. 26).

→ **Prototype selection**, e.g., prototype generation as a central element of the SVM+P technique proposed to generate explanations for SVM (Guidotti, et al., 2019, p. 26).

→ **Rules based explanations**, e.g., convert linear SVM-like model into a set of non-overlapping easily explainable rules (Guidotti, et al., 2019, p. 27), or extract if-then rules from previously identified prototypes with SVM+P (Guidotti, et al., 2019, p. 26).

→ **Visual techniques**, e.g., provide an explanation of SVM model as a colour-based nomogram representation (Belle, Calster, Huffel, Suykens, & Lisboa, 2016).

6.4 Usage and updating guidelines

[A12] Define the guidelines for using and updating the taxonomy. The constructed taxonomy is particularly valuable if used as intended, i.e., its usage is to be concretely specified. Furthermore, as the result of application of the constructed taxonomy together with further research in the underlying fields, the need to update the taxonomy will most likely emerge, hence the necessity to provide guidelines on appropriately updating the taxonomy.

The constructed taxonomy of explainability techniques of social media user profiling approaches in credit scoring can be used to support the process of selecting the best suitable social media user profiling approaches to derive specific consumers credit scoring model components under the consideration of the most appropriate explainability techniques available so to conform e.g., with law regulations, increase credit applicants' acceptance, etc. In other words, the constructed taxonomy provides paths from credit scoring model components through social media user profiling approaches to particular explainability techniques. Thus, the constructed taxonomy assists the decision-making process that requires expert knowledge in these three separate highly complex areas by interrelating them.

The updating of the constructed taxonomy is to proceed level-wise, since either not much research is expected to emerge on the overall issue of explainability techniques of social media user profiling approaches in credit scoring, but instead on single parts, which are (as earlier identified) credit scoring model components, social media user profiling approaches, and explainability techniques. Similar considerations apply to relations between single levels (i.e., the relational

structure of the taxonomy), which are to be updated separately: on the one side, the relations between credit scoring model components and social media user profiling approaches, and, on the other side, the relations between social media user profiling approaches and explainability techniques thereof.

7 Taxonomy validation

In the last phase of taxonomy development, the only one activity left to be performed is to validate the constructed taxonomy, **[A13] Validate the taxonomy**. Among the existing possibilities of taxonomy validation, for the constructed taxonomy it is decided to perform utility demonstration through classifying experts' opinion (Usman, Britto, Börstler, & Mendes, 2017, p. 45). Hence, research questions of the experts' opinion survey are derived from the respective research questions of the terms' extraction stage (conducted as SLRs) as follows.

CH7-RQ1. What are the components affecting the credit score of consumer credit applicants as per experts' opinion?

CH7-RQ2. What are the additional components potentially derivable from the social media data affecting the credit score of consumer credit applicants as per experts' opinion?

CH7-RQ3. What are the approaches for social media user profiling applicable to credit scoring as per experts' opinion?

CH7-RQ4. What are the techniques for explainability of social media user profiling approaches applicable to credit scoring as per experts' opinion?

7.1 Experts' opinion survey design

To successfully conduct the experts' opinion survey, as first, it is required to thoroughly define the survey research design, i.e., as a framework or a blue print (Sreejesh, 2014), detailing chosen survey research type, procedures for data collection, measurement, and analysis, survey research method, survey population and sampling.

7.1.1 Survey research type

The process of choosing the best suited overall type of survey research design is most often to make the decision between exploratory, descriptive, and causal survey research. (Sreejesh, 2014) endorses the necessity of being flexible, adaptable, efficient, and economy for a selected survey research design. At the same time, the most important requirement of all is naturally to result in an adequate information to solve posed survey research problems.

Explorative research helps to clarify vague research problems. It perfectly suits to be the very first step in a broad research initiative, supporting the process of enhancing research questions, formulating hypotheses, making first considerations regarding questionnaire structure, etc. (Sreejesh, 2014, p. 31). Hence, the results of the explorative study often serve as the source for the further research steps, which are either descriptive or causal research. These are then the two different types of conclusive research design, producing convincing results representative for a certain population. Taking into account the goal of validating the constructed taxonomy, precisely conclusive results are required. Moreover, the main issues to address have been already identified in the respective chapters.

Causal research, as the name suggests, aims to study cause and effect relationship between some particular research variables (Sreejesh, 2014, p. 82). In other words, it is aimed to study the influence of one variable on the change of another variable. (Sreejesh, 2014) further points out that the relationship can be either symmetric, asymmetric, or reciprocal. In order to conduct causal analysis, it is hence required to already possess an in-depth knowledge about the research problem. Thus, causal research is reasonable to conduct as the very last step of a more comprehensive research, i.e., after explorative and descriptive studies of a certain problem. Namely, having successfully identified research variables is one of the main prerequisites for starting causal analysis. (Burns, 2014) distinguish in this context between independent variables (have control and wish to manipulate them), dependent variables (little to no direct control, but strong interest in changing them), and extraneous variables (might have certain effect on dependable variables but are not yet independent). The main tool of causal research design is then the manipulation of independent variable with assessment of its influence on dependent variable, under the consideration of possible effects of additional extraneous variables, otherwise simply known as the process of experimentation (Burns, 2014, p. 79). Causal studies thus require thorough planning, they are very complicated to conduct, and as the result naturally expensive in terms of both required time and finances to execute them. This is another important reason of potentially starting with causal analysis only after other, much less expensive, types of research have been successfully done. In other words, there has to be a very strong primary conviction about the true need to conduct causal research.

Finally, descriptive research is the research design of choice for the desired survey to validate the constructed taxonomy. Similar to causal analysis, descriptive study proceeds precisely structured, producing results statistically inferable for the selected population. On the other hand, the focus of the descriptive research is on the defined population's single variable(s), instead of attempting

to identify cause and effect relationship between variables, as this is the case in causal research (Sreejesh, 2014) (Burns, 2014, p. 75). At the same time, it is not expected to receive unique insights from the outcome of a descriptive research, how it is otherwise the case by the explorative research, but rather to get an accurate description of the population of interest in terms of opinion, attitude, and other relevant characteristics regarding the research problem (Burns, 2014, p. 75). Validation of constructed taxonomy of explainability techniques of social media user profiling approaches in credit scoring by utility demonstration through classification of experts' opinion perfectly fits into this research design type.

The descriptive research studies are further classifiable into cross-sectional and longitudinal studies, depending on the time period over which they extend (Sreejesh, 2014, p. 60). Longitudinal studies are more seldom among them, running repeatedly over some period of time (Burns, 2014, p. 77). Some of the best-known longitudinal research designs include trend studies, pane studies, and cohort panels (Sreejesh, 2014, p. 61). They allow to measure the variation in the attitude, opinion, or other characteristics of the population regarding the research questions over a time period. Since that is not the aim of this work the posed by multiple measurements overhead is not justifiable. In contrast, cross-sectional studies are conducted at a single point in time, often termed as snapshot analysis (Burns, 2014, p. 75). This is sufficient, hence more preferable descriptive research design in the context of the desired survey.

7.1.2 Survey data collection

After making the precise choice on the survey research design type, it is important to further decide about the survey data collection mechanism. The main approaches in this context can be classified into qualitative, quantitative, or a mix thereof, depending on the nature of the research problem. Then, chosen data collection type respectively implies specific procedures involved during conducting the survey, such as those regarding sampling, data analysis and interpretation, etc.

Qualitative research involves collecting, analysing, and interpreting data largely by carefully observing and assessing what people do or say, etc. (Burns, 2014, p. 117) Free form or non-standardized nature of qualitative research with mostly open-ended questions allows to gather deeper and richer information from respondents (Burns, 2014, p. 138). Some of the most overall popular qualitative research techniques are observations, focus groups, and in-depth interviews, described by (Burns, 2014) as follows. *Observations* allow to observe how the respondents really behave, instead of relying on what the respondents think they would do. Their main disadvantage

observations share probably with all of the qualitative research techniques, namely usually relying on small samples, hence having limited representativeness. Another disadvantage is the subjective interpretation required to explain observed behaviour. *Focus groups* are usually moderated small group discussions with open communication focused on the research topic. Although they facilitate the generation of new ideas, focus groups are suggested to rather use for a better understanding of already present phenomena or to deepen into the findings from the quantitative studies. Again, focus groups have limited representativeness, require subjective evaluation of the discussions, in addition to high costs per participant. Finally, *in-depth interviews* provide the possibility to carefully examine motivation and considerations of participants regarding the research questions. At the same time, although in-depth interviews provide truly valuable possibility to gain access to participants' opinion, limited representativeness should be appropriately accounted for.

Quantitative research proceeds in a highly structured manner, operating based on a large number of respondents known to be representative for a wider population, hence removing the representativeness concern of qualitative studies at the cost of largely increased effort required for appropriate population sampling (Bryman, 2003, p. 11) (Burns, 2014, p. 118). The procedures of quantitative research types are very formalized, ordered, and numerical in nature, with clearly defined data gathering strategy largely based on closed-ended questions (Burns, 2014, p. 118).

As a side note, for a better complete understanding of the different data collection approaches it is useful to keep in mind that single studies are often not entirely either qualitative or quantitative, but rather contain some aspects and elements of each of those research data collection types. These so-called combined or mixed approaches further extend the otherwise understood continuum of research, i.e., interchangeable conducting of separate qualitative and quantitative studies. (Burns, 2014) describe it under the term of pluralistic research. Combining advantages of both qualitative and quantitative research takes place e.g., by beginning first with exploratory qualitative studies to build up a precise understanding of the problem or phenomenon prior to starting with the full-scale quantitative research. A qualitative phase hence serves as a foundation of a more expensive quantitative phase in the context of some comprehensive research project. Then, the chances are better for a more efficient and successful quantitative phase. In order to help the researchers to better understand the underlying motivation of the findings from the quantitative stage, subsequently qualitative research in some cases is again conducted (Burns, 2014, pp. 118-119).

The aim of the intended survey of this work is to gain a detailed view on the very specific topic of explainability techniques for social media user profiling approaches applicable to credit scoring. In other words, expert knowledge of credit scoring, social media user profiling, and explainability techniques is required. Hence, to keep the scope of this thesis manageable the decision is made against a quantitative survey, which is rather unfeasible in this context, and to decide in favour of a qualitative research as the most appropriate for the validation of the constructed taxonomy, with interviews and questionnaires selected as the preferred survey data collection method (Cresswell, 2012, pp. 217-222).

The last question to clarify regarding the source data for the research is the choice between primary and secondary data. Using primary data means to conduct the analysis based on the data gathered specifically for the underlying research project (Burns, 2014, p. 94). Secondary data has been gathered for other than the current research purposes, but is still additionally further utilizable (Burns, 2014, p. 94). Quick and comparably inexpensive access to data for the most different applications to be used as secondary data counts to its some of the most valuable advantages (Burns, 2014, p. 98). At the same time, since secondary data has not been collected to address specifically the same research questions as it is to be secondly used for, often such problems as an essential mismatch of measurement units, major differences in definitions, incompatible timeliness, and impossibility to confirm the credibility of data occur (Burns, 2014, p. 98). The novelty of the underlying research questions of this work together with stated disadvantages of secondary data essentially minimizes the possibilities of determining appropriate data to be used instead of collecting primary data. As usually, there is also a distinction between studies conducted exclusively based on primary or secondary data and research projects involving these both types of data. Secondary data might enhance primary data by providing first insights around the research questions, affecting what primary data would be then desired to be collected (Burns, 2014, p. 98). In other words, secondary data is often utilized in the explorative research. Hence, similar explanations as those provided in the research type subsection against conducting an explorative research also apply here to justify the needlessness of adopting secondary data in this work.

7.1.3 Population and sampling

To decide on the survey population, the relative novelty of the social media user profiling and of the explainability techniques of modern approaches were taken into account. At the same time, adoption of alternative data sources in the financial services industry is rather a slow process. As the result, two groups of experts are decided to be surveyed to answer the survey research

questions, since no single group of experts, most probably, would possess the knowledge to cover the fields of credit scoring, social media user profiling, and explainability techniques altogether.

- Expert group 1 (EG1): experts of credit scoring models and approaches for credit scoring, i.e., to address the survey research questions CH7-RQ1, CH7-RQ2, CH7-RQ3. The desired expertise and experience of the survey participants belonging to EG1 include at least 5 years of working experience in financial services industry with focus on consumers' credit scoring, in-depth knowledge and understanding of credit scoring models, their components and approaches to implement, knowledge of concepts, approaches, and possibilities of social media user profiling.
- Expert group 2 (EG2): experts of social media user profiling and explainability techniques, i.e., to address the survey research questions CH7-RQ3, CH7-RQ4. The desired expertise and experience of the survey participants belonging to EG2 include at least 5 years of working experience in the area of AI, data science, or similar (e.g., machine learning, statistical modelling, data analytics, etc.), in-depth knowledge and understanding of explainability techniques for various data-driven models, knowledge of concepts, approaches, and possibilities of social media user profiling.

The decision in favour of qualitative survey through interviews and questionnaires narrowed the choice of sampling approach to non-probabilistic sampling, namely the so-called purposeful sampling (Cresswell, 2012, p. 206). In purposeful sampling, individuals that can best help to learn or understand a certain topic are intentionally selected, choosing survey participants based on the information richness they can provide (Cresswell, 2012, p. 206). Several specific purposeful sampling strategies are available, differing in terms of being applied before or during data collection process, having different intent depending on the research problem and questions, etc., as follows (Cresswell, 2012, pp. 206-209).

- Maximal variation sampling: sample individuals that differ on particular characteristics or traits to present multiple perspectives from different groups.
- Extreme case sampling: study a particular outlier case or one that possesses certain extreme characteristics.
- Typical sampling: study individuals that are so-called typical to those unfamiliar with a particular situation.
- Theory or concept sampling: select individuals that can help to generate or to discover a theory or specific concepts within some theory.
- Homogenous sampling: sample individuals possessing similar traits or characteristics, i.e., belonging to a certain group of interest.

- Critical sampling: study individuals that represent central phenomenon of interest in dramatic (critical) terms.
- Opportunistic sampling: take advantage of events emerging after the study begins to gain additional insights.
- Snowball sampling: asking study participants to recommend other individuals to be sampled.
- Confirming and disconfirming sampling: strategy used during a study to follow up on particular cases to explore additional specific insights through confirming or disconfirming preliminary findings.

Since experts represent very specific groups that are aimed to be studied in-depth in order to validate the constructed taxonomy, hence homogenous sampling is selected to proceed with.

7.1.4 Interviews and questionnaire

Experts' opinion survey through interviews and questionnaire proceeds by asking general, open-ended questions that help to address defined research questions (Cresswell, 2012, p. 217). The questionnaire logically follows a possible discussion regarding explainability techniques for social media user profiling approaches in credit scoring. The main building blocks of the questionnaire respectively deal with credit scoring model components, social media user profiling approaches, and explainability techniques thereof. In order not to constrain survey participants' responses, open-ended questions are mainly asked (allowing participants to create their own options for responses), which are the most common in qualitative surveys (Cresswell, 2012, p. 218) The time required to answer of all of the questions was purposefully decided to keep in the range between 15 to 30 minutes in order to improve the response rate. For this sake, single questions are straight to the topic of the survey to efficiently address the survey research questions. The respective questionnaires are provided in Appendix O. Questionnaire of expert group 1 experts' opinion survey and Appendix P. Questionnaire of expert group 2 experts' opinion survey.

The most common single options to choose from regarding the exact approach of interviewing are: e-mail interviews, focus group interviews, one-on-one interviews, online questionnaire, etc. (Cresswell, 2012, pp. 217-222; Sreejesh, 2014, p. 62). E-mail interviews are useful to quickly collect data, e.g., from a geographically dispersed group of people (Cresswell, 2012, p. 219). Focus group interviews are very demanding to conduct and can be used to collect shared understanding from several individuals interviewed at the same time, hence requiring the

participants to be willing to cooperate with each other during the interview discussions (Cresswell, 2012, p. 218). One-on-one interviews are the most time-consuming, ideal for interviewing individuals who are not hesitant to speak and share ideas comfortably (Cresswell, 2012, p. 219). Online questionnaires are convenient for the survey participants, extremely time and cost efficient, with fast and reliable data collection, and eliminated interviewer bias (Van Selm, 2006, pp. 437-438). Hence, the decision was made to distribute the questionnaires to the experts by their choice through one of the following means: one-on-one interview, e-mail interview, online questionnaire. Thereby, the response rate is again increased by providing the possibility to choose the most convenient way to participate in the survey. At the same time, the process of extracting the experts' opinion from each of these three different types of questionnaire follows the same straightforward procedure, namely, the identification of mentioned terms of credit scoring model components, social media user profiling approaches, and explainability techniques in the respective experts' responses for the further classification of these terms with the constructed taxonomy for its validation. Hence, the survey conducted by different means is neither expected to negatively impact the ability to elicit the experts' opinion, nor creates the necessity for the additional comparison of the results to be performed afterwards.

7.2 Experts' opinion survey results

The survey questionnaire for the EG1 was presented to 13 experts, 5 responses were collected. The survey questionnaire for the EG2 was presented to 11 experts, 6 responses were collected. Results to defined research questions are presented in the following subsections.

7.2.1 Credit scoring model components

> CH7-RQ1. What are the components affecting the credit score of consumer credit applicants as per experts' opinion?

The following components that possibly affect the credit score of consumer credit applicants are elicited from the conducted experts' opinion survey: *a prior credit data, age, behaviour features, behaviour statistics, payment history, delinquency, credit bureau data, credit conditions specifics, credit amount, customer's income, demographic data, education, employment since, internal transactional records, marital status, net worth, occupation, property, savings history, small town vs. big city, transactional data, transactional features*.

The classification of the aforementioned components on the constructed taxonomy is as follows:

- Bank-borrower relationship: internal transactional records, savings history, transactional data, transactional features.
- Collateral characteristics: net worth, property, small town vs. big city.
- Credit applied for: a prior credit data, credit conditions specifics, credit amount.
- Credit card(s) data: –.
- Credit history: payment history, delinquency, credit bureau data.
- Demographic data: age, demographic data, education, marital status.
- Employment status: customer's income, employment since, occupation.
- Financial indicators: behaviour features, behaviour statistics.
- Look-a-likes: –.
- Psychological variables: –.
- Semiometric space: –.
- Social network data: –.
- User-generated content: –.

CH7-RQ2. What are the additional components potentially derivable from the social media data affecting the credit score of consumer credit applicants as per experts' opinion?

The following additional components potentially derivable from social media data that possibly affect the credit score of consumer credit applicants are elicited from the conducted experts' opinion survey: *demographic data, living above vs. below means, number of connections, number of posts, post behavioural, profiling re consumer behaviour, time frame within posts, type of connections, user posts*.

The classification of the aforementioned components on the constructed taxonomy is as follows:

- Bank-borrower relationship: ….
- Collateral characteristics: –.
- Credit applied for: –.
- Credit card(s) data: –.
- Credit history: –.
- Demographic data: *demographic data*.
- Employment status: –.
- Financial indicators: living above vs. below means, profiling re consumer behaviour.
- Look-a-likes: –.
- Psychological variables: –.

- Semiometric space: –.
- Social network data: number of connections, type of connections.
- User-generated content: number of posts, post behavioural, time frame within posts, user posts.

7.2.2 Social media user profiling approaches

CH7-RQ3. What are the approaches for social media user profiling applicable to credit scoring as per experts' opinion?

The following approaches for social media user profiling possibly applicable to credit scoring are elicited from the conducted experts' opinion survey: *Artificial Neural Networks, association rules, Bayesian methods, BERT, bi-LSTM, boosting models, clustering, CNN, Computer Vision Techniques, Convolutional Neural Network, Deep Learning techniques, Deep learning based, Deep models, deep neural networks, ensemble, Generalized additive model, hypothesis testing schemas, k-NN, k-means, likelihood maximization, linear regression, linear, logistics regression, Natural Language Processing, nearest neighbours, neural networks with stochastic component, neural networks, Probabilistic Graphical Models, stacked models, stacking learning, statistical inference-based approaches, SVM, tree models, tree-based, XGBoost*.

The classification of the aforementioned approaches on the constructed taxonomy is as follows:

- Artificial Neural Networks: Artificial Neural Networks, bi-LSTM, CNN, Computer Vision Techniques, Convolutional Neural Network, Deep Learning techniques, Deep learning based, Deep models, deep neural networks, neural networks with stochastic component, neural networks.
- Clustering: clustering, k-means.
- Decision trees: tree models, tree-based.
- Dimensionality reduction: –.
- Ensemble learning: boosting models, ensemble, stacked models, stacking learning, XGBoost.
- Graph theory algorithms: *association rules*.
- Linear models: Generalized additive model, linear regression, linear, logistics regression.
- Natural Language Processing: BERT, Natural Language Processing.
- Nearest neighbour models: *k-NN, nearest neighbours*.

- Probabilistic and statistical models: Bayesian methods, hypothesis testing schemas, likelihood maximization, Probabilistic Graphical Models, statistical inference-based approaches.
- Social recommender systems: –.
- Social semantic web: –.
- Support vector machines: *SVM*.

7.2.3 Explainability techniques for social media user profiling

CH7-RQ4. What are the techniques for explainability of social media user profiling approaches applicable to credit scoring as per experts' opinion?

The following techniques for explainability of social media user profiling approaches possibly applicable to credit scoring are elicited from the conducted experts' opinion survey: *Bayesian approach-based, Bayesian method, Deep Learning techniques, explainable surrogate models, glass box models, likelihood methods, LIME, linear-based models, Partial Dependency Plots, perturbation models, SHAP, SHAP values, statistical inference models, tailor models, tree-based methods, tree-based models, tree splits*.

The classification of the aforementioned techniques on the constructed taxonomy is as follows:

- Decision Tree based explanations: tree-based methods, tree-based models, tree splits.
- Deep explanations: Deep Learning techniques.
- Explainable surrogate models: LIME, explainable surrogate models.
- Features importance: *SHAP, SHAP values*.
- Model combination: ….
- Prototype selection: Bayesian method, Bayesian approach-based.
- Recommender systems explanations: ….
- Rules based explanations: ….
- Salient masks: ….
- Sensitivity analysis: perturbation models, likelihood methods, statistical inference models, tailor models.
- Textual justification: ….
- Transparent model types: glass box models, linear-based models.
- Visual techniques: Partial Dependency Plots.

8 Discussion

The taxonomy of explainability techniques for social media user profiling approaches applicable to credit scoring is constructed based on the knowledge identified from the conducted SLRs, and subsequently validated through classifying experts' opinion extracted from the conducted survey. Thorough discussions of the respective research questions RQ1 and RQ2 are provided in the both following subsections 8.1 and 8.2. Additionally, the subsection 8.3 provides the discussion of the experts' opinion on ethical issues resulting from the use of social media data in credit scoring.

The constructed taxonomy closely relates to the existing works that were identified in chapter 2. On the one hand, the general research of using social media for credit scoring (Ferretti, 2018; Guo, et al., 2016; Guo, et al., 2016; Eschholz, 2017) together with such specific research directions as legal issues of social media profiling (Ferretti, 2018; Eschholz, 2017; Goodman & Flaxman, 2017) are successfully extended by the view on the existing explainability techniques. On the other hand, the overall field of explainable AI (Hagras, 2018; Sheh & Monteath, 2018) together with explainability for such specific areas as decision support systems (Guidotti, et al., 2019; Nunes & Jannach, 2017; Waltl & Vogl, 2018) are concretized for the use case of social media user profiling approaches in credit scoring. Moreover, the limitation of lacking to comprehensively cover the concrete explainability techniques as the result of instead focusing on certain aspects of explainability, such as assessment of transparency (Waltl & Vogl, 2018), requirements to AI systems to be perceived as trustworthy (Sheh & Monteath, 2018), or other explanation aspects (Nunes & Jannach, 2017), is successfully addressed through the constructed taxonomy. Since none of the works identified in the state-of-the-art chapter are explicitly validated through an experts' opinion survey, the conducted validation of the constructed taxonomy of explainability techniques for social media user profiling approaches applicable to credit scoring is furthermore a major contribution.

Potential limitation of the constructed taxonomy results from the chosen source of knowledge for the conducted SLRs, which are only peer-reviewed academic publications. This decision, justified by the goal to contain only high-quality sources, leads to potentially missing knowledge contained in sources other than peer-reviewed publications. Future research can focus on additional sources of knowledge. Other specific limitations are mentioned in the following subsections where apply.

Overall, following the established methodology for the construction of taxonomies in software engineering, the comprehensive taxonomy is constructed in a structured and systematic manner. The constructed taxonomy is suitable to facilitate communication and application of the classified elements by researchers and practitioners through provided common terminology for knowledge sharing, improved understanding of the interrelationships, and identified gaps. Moreover, the taxonomy of explainability techniques for social media user profiling approaches in credit scoring supports the respective decision-making processes. The scientific developments on the field of explainability techniques are transferred to the field of social media user profiling approaches applicable in credit scoring, adapting the present knowledge for an adequately appropriate access. The constructed taxonomy is also extendable by researchers and practitioners upon emergence of additional credit scoring model components, development of new social media user profiling approaches or explainability techniques, and evidence for interrelations that are currently missing.

8.1 Taxonomy construction

The research question RQ1 is successfully addressed through the constructed taxonomy. Focus and consistency in handling the main three parts resulting from RQ1, which are credit scoring model components, social media user profiling approaches, and explainability techniques, as justified in chapter 5, proved to be very useful to successfully contain with the overall complexity. In the following subsections the specific discussion points regarding the categorization of the identified terms and the establishment of the relational structure are provided.

8.1.1 Taxonomy categories

The categorization of the identified terms was successfully conducted following the hybrid approach (Usman, 2015, p. 124) that combines traditional top-down and bottom-up approaches (Broughton, 2015). Namely, initial traversal of selected publications to extract explicitly mentioned categories, followed by conducting their terminology control, resulted in successful assignment of each of the extracted terms to a particular category.

8.1.2 Taxonomy relational structure

The relations between identified categories were successfully established through evaluation of identified terms. The few missing relational connections between individual categories are not a surprise and rather justified. On the one hand, for some of the credit scoring model components no evidence of existing social media user profiling approaches to derive them was found. Data on

bank-borrower relationship do not require to be explicitly derived from additional sources, and is utilizable as-is at the respective bank. Similar justification also applies to data on credit applied for, credit card(s) data, and credit history. Finally, user-generated content in the context of social media user profiling is part of the available data per se. On the other hand, for some of the social media user profiling approaches also no evidence of existing explainability techniques was found. The specific usage of dimensionality reduction mostly in conjunction with other approaches, in addition to rather straightforward idea behind it, possibly resulted in no explicit explainability techniques mentioned for it in the selected publications. Graph theory algorithms are much less common in the domain of AI in general, and machine learning in particular, than the other of the identified approaches, hence resulting in evidently no particular attention dedicated to study explainability of approaches of this specific category. Finally, ontologies, which are the underlying concept of social semantic web, are sometimes utilized to improve explainability of other approaches (Panigutti, Perotti, & Pedreschi, 2020) (Confalonieri, et al., 2019), being possibly the reason of no explicit explainability techniques for them found evident in the selected publications of the conducted SLR.

8.2 Taxonomy validation

The research question RQ2 is successfully addressed through conducted experts' opinion survey used to elicit experts' knowledge for its further classification. Experts from different regions of the world took part in the survey: from Africa (Nigeria), from Asia (India, Philippines, Turkey), from Europe (Austria, Denmark, France, Netherlands, Sweden), from South America (Argentina). Job positions held by the experts of the EG1 range from business analyst, business consultant, and business consulting manager to credit risk analyst, scoring analyst, statistical consultant, manager in credit and enterprise risk, and founder, fintech executive, data scientist, professor. Expertise in credit scoring, scorecards modelling, credit risk management, data science and data mining, machine learning, etc. is gained by the experts of the EG1 at various financial institutions (banks, credit reporting companies, fintech start-ups) and academic institutions (universities, research institutes). Experts of the EG2 have a strong academic background (5 of 6 with a Ph.D. degree). Job positions held by the experts of the EG2 range from researcher, data scientist, teaching and supervisory in the field of AI to technical consultant, technical architect, technical project manager, machine learning engineer. Expertise in data science, artificial intelligence, explainable AI, machine learning, etc. is gained by the experts of the EG2 at various academic institutions (universities, research institutes), consulting firms, start-ups, technology companies.

In the following subsections the specific discussion points regarding the taxonomy validation by experts' knowledge classification on each of the three main parts of the taxonomy are provided.

8.2.1 Credit scoring model components

The experts' knowledge regarding credit scoring model components is successfully classifiable by the created taxonomy, as shown in sections on results of the conducted experts' opinion survey. In addition to the core questions on credit scoring model components, the survey participants were also asked to provide feedback to the part of the constructed taxonomy that contain the categorization and the assignment of the single credit scoring model components' terms. Among 5 of the submitted ratings on the scale from 1 (worst score) to 5 (best score) the conducted credit scoring model components categorization received an average of 4.2 by the experts. Provided (optional) comments range from "*the list seems quite complete and covering the main areas (...)*" and "*it seems good enough for starting (...)*" to suggesting additional categories (e.g., macro-economic conditions, person's risk appetite and risk perception) and outlining own wholistic view on the challenges to identify the credit scoring model components in general or the potential of using specifically social media data for this purpose (e.g., different predictive power of different components, possible redundancy or less value-added, i.e., little importance, of social media data, reputation considerations, legal and ethical constraints).

8.2.2 Social media user profiling approaches

The experts' knowledge regarding social media user profiling approaches is successfully classifiable by the created taxonomy, as shown in section on results of the conducted experts' opinion survey. In addition to the core question on social media user profiling approaches, the survey participants were also asked to provide feedback to the part of the constructed taxonomy that contain the categorization and the assignment of the single social media user profiling approaches. Among 11 of the submitted ratings on the scale from 1 (worst score) to 5 (best score) the conducted social media user profiling approaches categorization received an average of 4.27 by the experts. Provided (optional) comments praise, in particular, the completeness of the conducted categorization ("*the list appears quite impressive (...)*", "*it seems very complete (...)*", etc.), though also pointing out possible difficulties in its understanding by the end-users caused by its complexity ("*(...) the literacy it demands to understand what is described is very high (...)*", etc.). Furthermore, there are also concrete suggestions for further approaches to be included, e.g., generalized additive model, independent component analysis, non-negative matrix factorization.

Finally, the advantage of the EG1 participating experts of possessing knowledge on credit scoring model components and social media user profiling approaches is utilized with an additional question, namely to provide feedback to the relational part of the constructed taxonomy that connect particular social media user profiling approaches potentially applicable to derive particular credit scoring model components as by made categorizations. Among 5 of the submitted ratings on the scale from 1 (worst score) to 5 (best score) the aforementioned relational part of the constructed taxonomy received an average of 4.2 by the experts. Sounded critical points address predominately theoretical nature of the elaborated connections between credit scoring model components and social media user profiling approaches possibly applicable to derive them, suggesting to conduct practical evaluation in the future work, and the overall potential difficulty to follow the many-to-many relationships as expressed in the conducted taxonomy, what is accounted for through depicting relationships of single credit scoring model components at a time in the respective subsections on the taxonomy relational structure.

8.2.3 Explainability techniques for social media user profiling

The experts' knowledge regarding explainability techniques is successfully classifiable by the created taxonomy, as shown in section on results of the conducted experts' opinion survey. In addition to the core question on explainability techniques, the survey participants were also asked to provide feedback to the part of the constructed taxonomy that contain the categorization and the assignment of the single explainability techniques. Among 6 of the submitted ratings on the scale from 1 (worst score) to 5 (best score) the conducted explainability techniques categorization received an average of 4.17 by the experts. Provided (optional) comments range from "*nice list*" and praising the completeness of the conducted categorization ("*(...) the categorization seems very complete (...)*", etc.) to pointing out the potential difficulties in its understanding caused by its complexity ("*(...) not written in a form that allows for the layman to understand (...)*", etc.) or possible contextual issues regarding applicability for the credit scoring, which are accounted for through depicting relationships of single social media user profiling approaches at a time in the respective subsections on the taxonomy relational structure and an extra level of credit scoring model components as contextual information of applicability to credit scoring.

Finally, the advantage of the EG2 participating experts of possessing knowledge on social media user profiling approaches and explainability techniques is utilized with an additional question, namely to provide feedback to the relational part of the constructed taxonomy that connect particular explainability techniques potentially applicable to explain particular social media user profiling approaches as by conducted categorizations. Among 6 of the submitted ratings on the

scale from 1 (worst score) to 5 (best score) the aforementioned relational part of the constructed taxonomy received an average of 4.17 by the experts. Sounded critical points suggest more connections could be added, i.e., further concrete explainability techniques potentially applicable to explain certain categories of approaches, if not the limitation to social media user profiling approaches would be in place, and the potential weakness of expressing the relational structure in the elaborated visual manner for the overall understandability, what is accounted for through depicting relationships of single social media user profiling approaches at a time in the respective subsections on the taxonomy relational structure.

8.3 Experts' opinion on ethical issues of social media data in credit scoring

As a follow-up to the conducted experts' opinion survey for the validation of the constructed taxonomy, the decision was made to additionally elicit experts' opinion on possible ethical issues resulting from using social media data in credit scoring. The single asked question is as follows: what ethical concerns and potential issues resulting from using social media user profiling for credit scoring would you identify? The experts' feedback is discussed in the following.

Identified potential ethical issues with using social media data in credit scoring repeatedly acknowledge the general controversy behind such undertaking. Provided feedback ranges from expressed concerns for "*growth of power the social networks will gain to govern people*" to even fear of thereby created potential to "*curtail freedom of expression and right to protest*".

One of the most often mentioned concerns is related to discrimination, arguing that utilizing social media for credit scoring could facilitate some sort of discrimination, such as racial, economic, religious, etc. Deriving additional sensitive information from social media about the credit applicant, not explicitly provided for the purpose of credit scoring, might be "*yielding high discriminatory power*" besides potentially even "*not be legally usable*". This re-iterates both major issues of ethics of data as examined in the subsection 3.3.1.

The ethical issues identified by the experts in regard to possible underlying algorithms are expressed the most in-depth, providing various potential examples. As experts argue, "*most of the social media content is contextual and views are relative*". Hence, an algorithm conducting social media user profiling could perform unethically, unfairly inferring "*from a short text to ascertain a view or potential action of a customer*" or assuming "*that social media connections of an*

applicant indirectly hint at his/her ability to repay a loan". A major ethical issue arises from the consideration that "*dishonest applicants can create fake profiles*", thus, possibly harming the system's performance, potentially also negatively impacting other applicants. Furthermore, taking into account that "*algorithms may assign a new applicant to a group of existing borrowers, based on similarity of profiles, then any mistake (either false positive or false negative)*" could similarly lead to an impaired system's performance, or, for example, "*an innocent applicant may be suspected in fraud if his profile is close enough to profiles of typical fraudsters that a credit organization has previously collected*". In case of faulty evidence stored and propagated to other databases, it might be not possible anymore to "*guarantee the same change in other databases*" once that mistake is identified and corrected in the original system. These ethical concerns re-iterate major issues of ethics of algorithms as examined in the subsection 3.3.2.

Some experts question principles under which data would potentially be shared with external parties, and the overall concern of information published on social media not intended for credit scoring purposes being used for it. The very specific ethical issue of requirements on explicit and informed consent for social media profiling conducted for credit scoring concerns some other experts, in particular the necessity "*to let the customers know which aspects of social media will be picked and how are each scored* ", and that "*the user should be let known of what exactly is being profiled*". Furthermore, since "*giving access to own social media profile certainly undermines applicants' security and privacy*" there is also a certain pressure on those in need, i.e., "*desperate borrowers*". These ethical concerns re-iterate major issues of ethics of practices as examined in the subsection 3.3.3.

Summing up, the experts' opinion highly correlates with the overall potential issues elaborated in the subsection 3.3. Concluding the discussion of ethical concerns resulting from the potential undertaking of using social media for credit scoring, one expert expressed the opinion that "*scoring models should be interpretable or explainable as much as possible in order to detect any bias and to timely correct it*" to mitigate at least to certain extent possible unethical impact.

9 Conclusion

The aim of this thesis to construct taxonomy of explainability techniques for social media user profiling approaches potentially applicable to derive credit scoring model components was successfully achieved following the methodology for developing taxonomies in software engineering (Usman, Britto, Börstler, & Mendes, 2017). In particular, systematic literature reviews (Kitchenham & Charters, 2007) are extensively applied for taxonomy terms identification followed by their categorization through a hybrid approach (Usman, 2015, p. 124) that combines traditional top-down and bottom-up approaches (Broughton, 2015), establishing respective relational structure between identified categories. The validation of the taxonomy is successfully accomplished through classifying the experts' opinion extracted from the conducted experts' opinion survey. Moreover, the necessity to account also for potential ethical issues of such rather controversial undertaking as utilizing social media user profiling for credit scoring is successfully addressed by the respective in-depth discussion that followed the approach to assess the ethical implications of data science elaborated by (Floridi & Taddeo, 2016).

Selected methodological approach proved to be useful to elicit respective terms regarding credit scoring model components (selected 40 publications analysed in-depth), social media user profiling approaches (selected 85 publications analysed in-depth), and explainability techniques (selected 33 publications analysed in-depth) from high quality sources, which the total of 158 selected and analysed in-depth peer-reviewed academic publications are. Similarly, relational structure of the constructed taxonomy is also entirely based on the strong evidence extracted from the same sources of high quality. The resulting possible limitation is lack of terms and relations potentially mentioned in other sources of lesser quality.

The constructed taxonomy is an important tool for theory and practice: on the one side, researchers that propose either additional credit scoring model components or new social media user profiling approaches capable to potentially be used in credit scoring can easily align with the developed taxonomy particularly regarding existence of suitable explainability techniques thereof, and, on the other side, practitioners in credit scoring domain can easily incorporate the developed taxonomy in their decision making process on utilizing either additional credit scoring model components or new social media user profiling approaches regarding existence of suitable explainability techniques thereof. Consequently, future research could address the components

currently not evident to be derivable from social media data by any of the social media user profiling techniques, and social media user profiling approaches for which currently no evidence for the existence of explainability techniques available. Furthermore, as the experts that participated in the validation of the constructed taxonomy confirm, another important future research direction is to constantly preserve the practical applicability of the overall taxonomy.

This work contributed a novel comprehensive view of the explainability techniques for social media user profiling approaches potentially applicable to credit scoring in a systematic and structured manner through categorization of the respective terms of credit scoring model components, social media user profiling approaches, and explainability techniques thereof, and establishing the relations between identified categories. In other words, the problem of which are the techniques to provide explainability to social media profiling in credit scoring is addressed through the developed extensive taxonomy, which is also successfully validated through classifying experts' opinion extracted from conducted experts' opinion survey. Finally, the elaborated comprehensive view on possible ethical issues of utilizing social media user profiling for credit scoring strongly emphasizes the necessity to always appropriately consider potential ethical implications specifically for all novel undertakings in the field of data science.

References

Abbod, M., & Radi, M. (2018). The applicability of credit scoring models in emerging economies: an evidence from Jordan. *International Journal of Islamic and Middle Eastern Finance and Management, Vol.11(4)*, 608-630.

Abdou, H. A. (2009). An evaluation of alternative scoring models in private banking. *The Journal of Risk Finance, Vol.10(1)*, 38-53.

Abdou, H., Alam, S., & Mulkeen, J. (2014). Would credit scoring work for Islamic finance? A neural network approach. *International Journal of Islamic and Middle Eastern Finance and Management, Vol.7(1)*, 112-125.

Abeer, H., & B, H. W. (2016). Credit Risk Assessment Model Based Using Principal component Analysis And Artificial Neural Network. *MATEC Web of Conferences, Vol.76*, 02039.

Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2016). The Consumer Loan's Payment Default Predictive Model: An Application in a Tunisian Commercial Bank. *Asian Economic and Financial Review, Vol.6(1)*, 27-42.

Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2018). The Consumer Loan's Payment Default Predictive Model: an Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank. *Journal of the Knowledge Economy, Vol.9(3)*, 948-962.

Abid, L., Zaghdene, S., & Masmoudi, A. (2017). Bayesian Network Modeling: A Case Study of Credit Scoring Analysis of Consumer Loans Default Payment. *Asian Economic and Financial Review, Vol.7(9)*, 846-857.

About Lenddo. (2018, 01 25). Retrieved from Leveraging Technology Solutions in Credit and Verification | Lenddo: https://www.lenddo.com/about.html

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research, Vol.222(1)*, 168-178.

Al Hasan Haldar, N., Li, J., Reynolds, M., Sellis, T., & Yu, J. (2019). Location prediction in large-scale social networks: an in-depth benchmarking study. *The VLDB Journal, Vol.28(5)*, 623-648.

Al-Daraiseh, A. A., Al-Joudi, A. S., Al-Gahtani, H. B., & Al-Qahtani, M. S. (2014). Social networks' benefits, privacy, and identity theft: KSA case study. *Social Networks, 5(12)*, pp. 129-143.

Ali, S., Islam, N., Rauf, A., Din, I., Guizani, M., & Rodrigues, J. (2018). Privacy and Security Issues in Online Social Networks. *Future Internet, 10(12)*, p. 114.

Alibaba: cumulative active online buyers. (2018, 02 01). Retrieved from Statista: https://www.statista.com/statistics/226927/alibaba-cumulative-active-online-buyers-taobao-tmall/

Alipay. (2018, 02 01). Retrieved from https://intl.alipay.com/

Alpar, P. (2016). Offshoring in the Wrong Direction? *International Workshop on Global Sourcing of Information Technology and Business Processes* (pp. 166-177). Springer, Cham.

Al-Qurishi, M., Alhuzami, S., AlRubaian, M., Hossain, M. S., Alamri, A., & Rahman, M. (2018). User profiling for big social media data using standing ovation model. *Multimedia Tools and Applications, Vol.77(9)*, 11179-11201.

Alshammari, A., Kapetanakis, S., Polatidis, N., Evans, R., & Alshammari, G. (2019). Twitter user modeling based on indirect explicit relationships for personalized recommendations. *International Conference on Computational Collective Intelligence* (pp. 93-105). Springer, Cham.

Alshammari, M., Nasraoui, O., & Sanders, S. (2019). Mining Semantic Knowledge Graphs to Add Explainability to Black Box Recommender Systems. *IEEE Access, Vol.7*, 110563-110579.

Amal, S., Tsai, C.-H., Brusilovsky, P., Kuflik, T., & Minkov, E. (2019). Relational social recommendation: Application to the academic domain. *Expert Systems With Applications, Vol.124*, 182-195.

Anand, D., & Mampilli, B. (2014). User profiling based on keyword clusters for improved recommendations. *International Conference on Distributed Computing and Internet Technology* (pp. 176-187). Springer, Cham.

Anand, D., & Mampilli, B. S. (2014). Folksonomy-based fuzzy user profiling for improved recommendations. *Expert Systems With Applications, Vol.41(5)*, 2424-2436.

Anderson, B. (2019). Using Bayesian networks to perform reject inference. *Expert Systems With Applications, Vol.137*, 349-356.

Arain, Q. A., Memon, H., Memon, I., Memon, M. H., Shaikh, R. A., & Mangi, F. A. (2017). Intelligent travel information platform based on location base services to predict user travel behavior from user-generated GPS traces. *International Journal of Computers and Applications, Vol.39(3)*, 155-168.

Ariza, M., Arroyo, J., Caparrini, A., & Segovia, M.-J. (2020). Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access*.

Arras, L., Horn, F., Montavon, G., Müller, K.-R., & Samek, W. (2017). "What is relevant in a text document?": An interpretable machine learning approach. *PloS one, Vol.12(8)*, e0181142.

Arun, R. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, Vol.48(1)*, 137-141.

Backer, L. C. (2017). Measurement, Assessment and Reward: The Challenges of Building Institutionalized Social Credit and Rating Systems in China and in the West. *Proceedings of the Chinese Social Credit System.* Shanghai Jaiotong University.

Baklouti, I. (2014a). A psychological approach to microfinance credit scoring via a classification and regression tree. *Intelligent Systems in Accounting, Finance and Management, Vol.21(4)*, 193.

Bakshy, E., Dean, E., Rong, Y., & Itamar, R. (2012). Social influence in social advertising: evidence from field experiments. *Proceedings of the 13th ACM Conference on Electronic Commerce*, 146-161.

Banasik, J., & Crook, J. (2005). Credit scoring, augmentation and lean models. *Journal of the Operational Research Society: Special Issue: Credit Scoring, Vol.56(9)*, 1072-1081.

Barbon, S., Igawa, R., & Bogaz Zarpelão, B. (2017). Authorship verification applied to detection of compromised accounts on online social networks. *Multimedia Tools and Applications, Vol.76(3)*, 3213-3233.

Barysheva, A., Petrov, M., & Yavorskiy, R. (2015). Building profiles of blog users based on comment graph analysis: The Habrahabr.ru case. *Communications in Computer and Information Science, Vol.542*, 257-262.

Béjar, J., Álvarez, S., García, D., Gómez, I., Oliva, L., Tejeda, A., & Vázquez-Salceda, J. (2016). Discovery of spatio-temporal patterns from location-based social networks. *Journal of Experimental & Theoretical Artificial Intelligence, Vol.28(1-2)*, 313-329.

Belle, V. V., Calster, B. V., Huffel, S. V., Suykens, J. A., & Lisboa, P. (2016). Explaining Support Vector Machines: A Color Based Nomogram. *PloS one, Vol.11(10)*, e0164568.

Bennacer Seghouani, N., Jipmo, C., & Quercini, G. (2019). Determining the interests of social media users: two approaches. *Information Retrieval Journal, Vol.22(1)*, 129-158.

Besel, C., Schlötterer, J., & Granitzer, M. (2016). On the quality of semantic interest profiles for onine social network consumers. *ACM SIGAPP Applied Computing Review, Vol.16(3)*, 5-14.

Bharadhwaj, H., & Joshi, S. (2018). Explanations for Temporal Recommendations. *KI - Künstliche Intelligenz, Vol.32(4)*, 267-272.

Bikmukhametov, T., & Jäschke, J. (2020). Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Computers and Chemical Engineering, Vol.138*.

Bilge, L., Strufe, T., Balzarotti, D., & Kirda, E. (2009). All your contacts are belong to us: automated identity theft attacks on social networks. *In Proceedings of the 18th international conference on World wide web* (pp. 551-560). ACM.

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. *Frontiers in aging neuroscience, Vol.11*, 194.

Bok, K., Yoon, S., & Yoo, J. (2019). Trust evaluation of multimedia documents based on extended provenance model in social semantic web. *Multimedia Tools and Applications, 78*, 28681–28702.

Bourque, P., & Fairley, R. E. (2014). *Guide to the Software Engineering Body of Knowledge, Version 3.0.* IEEE Computer Society.

Boyd, D. M., & Ellison, N. B. (2010). Social network sites: definition, history, and scholarship. *IEEE Engineering Management Review, 38(3)*, 16-31.

Broughton, V. (2015). *Essential classification.* Facet Publishing.

Bryman, A. (2003). *Quantity and quality in social research.* Routledge.

Buraya, K., Farseev, A., & Filchenkov, A. (2018). Multi-view personality profiling based on longitudinal data. *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 15-27). Springer, Cham.

Burns, A. C. (2014). *Marketing research (Vol. 7).* Harlow: Pearson.

Bynum, T. (2015). *Computer and information ethics*. Retrieved from Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/win2015/entries/ethics-computer/

C C, N., & Mohan, A. (2019). A social recommender system using deep architecture and network embedding. *Applied Intelligence, Vol.49(5)*, 1937-1953.

Cady, F. (2017). *The Data Science Handbook.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics (Basel), Vol.8(8)*, 832.

Ceni, A., Ashwin, P., & Livi, L. (2020). Interpreting Recurrent Neural Networks Behaviour via Excitable Network Attractors. *Cognitive Computation, Vol.12(2)*, 330-356.

Chen, D., Zhang, Q., Chen, G., Fan, C., & Gao, Q. (2018). Forum User Profiling by Incorporating User Behavior and Social Network Connections. *International Conference on Cognitive Computing* (pp. 30-42). Springer, Cham.

Chen, E., Zeng, G., Luo, P., Zhu, H., Tian, J., & Xiong, H. (2017). Discerning individual interests and shared interests for social user profiling. *World Wide Web, Vol.20(2)*, 417-435.

Chen, L., Yan, D., & Wang, F. (2019). User Evaluations on Sentiment-based Recommendation Explanations. *ACM Transactions on Interactive Intelligent Systems (TiiS), Vol.9(4)*, 1-38.

Chen, X., Wang, J., Ren, Y., Liu, T., & Lin, H. (2018). NLPCC 2018 Shared Task User Profiling and Recommendation Method Summary by DUTIR_9148. *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 420-428). Springer, Cham.

Chen, Y., & Cheung, A. S. (2017). The Transparent Self Under Big Data Profiling: Privacy and Chinese Legislation on the Social Credit System. *The University of Hong Kong, Faculty of Law, Legal Studies Research Paper Series, 7(4)*.

Chen, Y., & Cheung, A. S. (2017). The Transparent Self Under Big Data Profiling: Privacy and Chinese Legislation on the Social Credit System (No. 2017/011). *The University of Hong Kong, Faculty of Law, Legal Studies Research Paper Series, 7(4)*.

Chen, Z., Zhu, F., Guo, G., & Liu, H. (2014). User profiling via affinity-aware friendship network. *International Conference on Social Informatics* (pp. 151-165). Springer, Cham.

Cheng, C., & Ou, S. Y. (2014). The status quo and problems of the building of china's social credit system and suggestions. *International Business and Management, 8(2)*, pp. 169-173.

Cheng, Z., Chang, X., Zhu, L., Kanjirathinkal, R., & Kankanhalli, M. (2019). MMALFM: Explainable Recommendation by Leveraging Reviews and Images. *ACM Transactions on Information Systems (TOIS), Vol.37(2)*, 1-28.

Chi, B.-W., & Hsu, C.-C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems With Applications, Vol.39(3)*, 2650-2661.

Chuang, C.-L., & Lin, R.-H. (2009). Constructing a reassigning credit scoring model. *Expert Systems With Applications, Vol.36(2)*, 1685-1694.

Confalonieri, R., del Prado, F. M., Agramunt, S., Malagarriga, D., Faggion, D., Weyde, T., & Besold, T. R. (2019). An Ontology-based Approach to Explaining Artificial Neural Networks. *arXiv preprint arXiv:1906.08362*.

Costa, A., Deb, A., & Kubzansky, M. (2015). Big data, small credit: The digital revolution and its impact on emerging market consumers. *Innovations: Technology, Governance, Globalization, 10(3-4)*, pp. 49-80.

Crawford, K., & Finn, M. (2015). The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal, 80(4)*, 491-502.

Creemers, R. (2015). Planning Outline for the Construction of a Social Credit System (2014-2020).

Cresswell, J. W. (2012). *Planning, conducting, and evaluating quantitative and qualitative research.* Educational Research.

Crook, J., Hamilton, R., & Thomas, L. (1992). A Comparison of a Credit Scoring Model with a Credit Performance Model. *The Service Industries Journal, Vol.12(4)*, 558-579.

Cullerton, N. (2012). Behavioral credit scoring. *Geo. LJ, 101*, 807.

Datenschutzgesetz. (2020). Retrieved from Rechtsinformationssystem des Bundes: https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=bundesnormen&Gesetzesnummer=10001597

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting Depression via Social Media. *Seventh international AAAI conference on weblogs and social media*.

De Cnudde, S., Moeyersoms, J., Stankova, M., Tobback, E., Javaly, V., & Martens, D. (2019). What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance. *Journal of the Operational Research Society, Vol.70(3)*, 353-363.

De Koninck, P., De Weerdt, J., & vanden Broucke, S. (2017). Explaining clusterings of process instances. *Data Mining and Knowledge Discovery, Vol.31(3)*, 774-808.

De Salve, A., Guidi, B., Ricci, L., & Mori, P. (2018). Discovering Homophily in Online Social Networks. *Mobile Networks and Applications, Vol.23(6)*, 1715-1726.

Definition of Ethics by Oxford Dictionary. (2021, 01 31). Retrieved from https://www.lexico.com/definition/ethics

Deville, J. (2013). Leaky data: How Wonga makes lending decisions. *Charisma: Consumer Market Studies*.

Dharia, S., Eirinaki, M., Jain, V., Patel, J., Varlamis, I., Vora, J., & Yamauchi, R. (2018). Social recommendations for personalized fitness assistance. *Personal and Ubiquitous Computing, Vol.22(2)*, 245-257.

di Vimercati, S. D., Foresti, S., & Samarati, P. (2012). Managing and accessing data in the cloud: Privacy risks and approaches. *2012 7th International Conference on Risks and Security of Internet and Systems (CRiSIS)* (pp. 1-9). IEEE.

Diab, R. S. (2017). Becoming-Infrastructure: Datafication, Deactivation and the Social Credit System. *Journal of Critical Library and Information Studies, 1(1)*.

Dimitriu, M., Avramescu, E. A., & Caracota, R. C. (2010). Credit scoring for individuals. *Economia: Seria Management, Vol.13(2)*, 361-377.

Dinev, T., Hart, P., & Mullen, M. R. (2008). Internet privacy concerns and beliefs about government surveillance–An empirical investigation. *he Journal of Strategic Information Systems, 17(3)*, pp. 214-223.

Dougnon, R., Fournier-Viger, P., Lin, J., & Nkambou, R. (2016). Inferring social network user profiles using a partial social graph. *Journal of Intelligent Information Systems, Vol.47(2)*, 313-344.

Dougnon, R., Fournier-Viger, P., Lin, J.-W., & Nkambou, R. (2015). Accurate online social network user profiling. *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)* (pp. 264-270). Springer, Cham.

Drmota, M., Gittenberger, B., Karigl, G., & Panholzer, A. (2007). *Mathematik für Informatik.* Berlin: Heldermann Verlag.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM, Vol.63(1)*, 68-77.

Ebnet, N. J. (2012). It can do more than protect your credit score: Regulating social media pre-employment screening with the fair credit reporting act. *Minnesota Law Review, 97*, 306.

Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F. (2019). A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. *IEEE Access, Vol.7*, 144907-144924.

Eschholz, S. (2017). Big Data-Scoring unter dem Einfluss der Datenschutz-Grundverordnung. *Datenschutz und Datensicherheit-DuD, 41(3)*, 180-185.

Eyharabide, V., & Amandi, A. (2012). Ontology-based user profile learning. *Applied Intelligence, Vol.36(4)*, 857-869.

Fang, Q., Sang, J., Xu, C., & Hossain, M. S. (2015). Relational User Attribute Inference in Social Media. *IEEE Transactions on Multimedia, Vol.17(7)*, 1031-1044.

Faralli, S., Stilo, G., & Velardi, P. (2015). Recommendation of microblog users based on hierarchical interest profiles. *Social Network Analysis and Mining, Vol.5(1)*, 1-23.

Ferretti, F. (2018). Not-So-Big and Big Credit Data Between Traditional Consumer Finance, FinTechs, and the Banking Union: Old and New Challenges in an Enduring EU Policy and Legal Conundrum. *Global Jurist, 18(1)*.

Floridi, L., & Taddeo, M. (2016). What is Data Ethics? *Philosophical Transactions of the Royal Society A, Volume 374, Issue 2083*.

França, F., Goya, D., & Camargo Penteado, C. (2018). User profiling of the Twitter Social Network during the impeachment of Brazilian President. *Social Network Analysis and Mining, Vol.8(1)*, 1-10.

Friedrich, A. (2018, 01 25). *Traditional Lenders Nurture a Two-Class System*. Retrieved from Kreditech: https://www.kreditech.com/magazine/traditional-lenders-nurture-two-class-system/

Gafny, M. A., Shabtai, A., Rokach, L., & Elovici, Y. (2010). Detecting data misuse by applying context-based data linkage. *Proceedings of the 2010 ACM workshop on Insider threats* (pp. 3-12). ACM.

Gan, C., Li, Z., Wang, W., & Kao, B. (2012). Credit scoring in mortgage lending: evidence from China. *International Journal of Housing Markets and Analysis, Vol.5(4)*, 334-350.

Gao, H., Barbier, G., & Goolsby , R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems, 26(3)*, 10-14.

Gao, H., Hu, J., Huang, T., Wang, J., & Chen, Y. (2011). Security issues in online social networks. *IEEE Internet Computing, 15(4)*, pp. 56-63.

Gehrlein, W., & Wagner, B. (1997). A two-stage least cost credit scoring model. *Annals of Operations Research, Vol.74*, 159-171.

Global social media ranking 2019 | Statista. (2020). Retrieved from Statista: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI magazine, 38(3)*, 50-57.

Gorrab, A., Kboubi, F., Jaffal, A., Le Grand, B., & Ghezala, H. (2017). Twitter user profiling model based on temporal analysis of hashtags and social interactions. *International Conference on Applications of Natural Language to Information Systems* (pp. 124-130). Springer, Cham.

Gu, X., Yang, H., Tang, J., Zhang, J., Zhang, F., Liu, D., . . . Fu, X. (2018). Profiling Web users using big data. *Social Network Analysis and Mining, Vol.8(1)*, 1-17.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR), 51(5)*, 1-42.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR), Vol.51(5)*, 1-42.

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence Program. *AI Magazine, Vol.40(2)*, 44-58.

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. *ACM Transactions on the Web (TWEB), Vol.10(4)*, 1-38.

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. *ACM Transactions on the Web (TWEB), Vol.10(4)*, 1-38.

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. *ACM Transactions on the Web (TWEB), 10(4)*, 1-38.

Guo, G., Zhu, F., Chen, E., Wu, L., Liu, Q., Liu, Y., & Qiu, M. (2016). Personal credit profiling via latent user behavior dimensions on social media. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 130-142). Springer, Cham.

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics, 118(1)*, pp. 177-214.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and machines (Dordrecht), Vol.30 (1)*, pp. 99-120.

Hagras, H. (2018). Toward Human-Understandable, Explainable AI. *Computer, 51(9)*, 28-36.

Han, D. (2017). The Market Value of Who We Are: The Flow of Personal Data and Its Regulation in China. *Media and Communication, 5(2)*, (pp. 21-30).

Harpring, P. (2010). *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works.* Getty Publications.

Hirt, R., Kühl, N., & Satzger, G. (2019). Cognitive computing for customer profiling: meta classification for gender prediction. *Electronic Markets, Vol.29(1)*, 93-106.

Hoang, T., & Lim, E. (2017). Modeling Topics and Behavior of Microbloggers: An Integrated Approach. *Acm Transactions On Intelligent Systems And Technology, Vol.8(3)*.

Hong, M., Akerkar, R., & Jung, J. J. (2019). Improving Explainability of Recommendation System by Multi-sided Tensor Factorization. *Cybernetics and Systems: adding smartness to systems with case studies and applications, Vol.50(2), pp.97-117*, 97-117.

Hou, Y., Yang, N., Wu, Y., & Yu, P. (2019). Explainable recommendation with fusion of aspect information. *World Wide Web, Vol.22(1)*, 221-240.

Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems With Applications, Vol.28(4)*, 655-665.

Huang, X., Yang, Z., Yang, Y., Shen, F., Xie, N., & Shen, H. (2017). A deep approach for multi-modal user attribute modeling. *Australasian Database Conference* (pp. 217-230). Springer, Cham.

Huang, Y., Yu, L., Wang, X., & Cui, B. (2015). A multi-source integration framework for user occupation inference in social media systems. *World Wide Web, Vol.18(5)*, 1247-1267.

Huang, Z., Lei, Y., & Shen, S. (2016). China's personal credit reporting system in the internet finance era: challenges and opportunities. *China Economic Journal, 9(3)*, pp. 288-303.

Huch, S. (2016). Fallbeispiele innovativer Fintech-Unternehmen. *Wirtschaftsinformatik & Management, 8(3)*, pp. 64-73.

Hunter, R. F., Gough, A., O'Kane, N., McKeown, G., Fitzpatrick, A., Walker, T., . . . Kee, F. (2018). Ethical issues in social media research for public health. *American journal of public health 108, no. 3*, 343-348.

Janeska, M., Taleska, S., & Sotiroski, K. (2014). Application of the Scoring Model for Assessing the Credit Rating of Principals. *TEM Journal, Vol.3(1)*, 50-54.

Jansen, B., Jung, S.-G., Salminen, J., An, J., & Kwak, H. (2018). Combining behaviors and demographics to segment online audiences: Experiments with a youtube channel. *International Conference on Internet Science* (pp. 141-153). Springer, Cham.

Ju, C., & Tao, W. (2017). A novel relationship strength model for online social networks. *Multimedia Tools and Applications, Vol.76(16)*, 17577-17594.

Käde, L., & Von Maltzan, S. (2019). Towards a demystification of the Black Box – explainable AI and legal ramifications. *Journal of Internet Law, Vol.23(3)*, 3-13.

Kandias, M., Mitrou, L., Stavrou, V., & Gritzalis, D. (2014). Youtube user and usage profiling: Stories of political horror and security success. *Communications in Computer and Information Science, Vol.456*, 270-289.

Kang, J., Choi, H., & Lee, H. (2019). Deep recurrent convolutional networks for inferring user interests from social media. *Journal of Intelligent Information Systems, Vol.52(1)*, 191-209.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53(1)*, (pp. 59-68).

Karlsson, I., Rebane, J., Papapetrou, P., & Gionis, A. (2019). Band Selection via Explanations From Convolutional Neural Networks. *Knowledge and Information Systems*.

Kim, Y. S., & Sohn, S. Y. (2004). Managing loan customers using misclassification patterns of credit scoring model. *Expert Systems With Applications, Vol.26(4)*, 567-573.

Kiss, F. (2003). Credit scoring processes from a knowledge management perspective. *Periodica Polytechnica. Social and Management Sciences, Vol.11(1)*, 95-110.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering (Vol. 5).* Technical report, Ver. 2.3. EBSE.

Koh, H., & Tan, W. (2004). Credit Scoring Using Data Mining Techniques. *Singapore Management Review, Vol.26(2)*, 25-47.

Komesaroff, P. A., & Parker, M. (2009). Ethical Aspects of Consent. *Issues 86*, p. 24.

Kshetri, N. (2016). Big data's role in expanding access to financial services in China. *International journal of information management, 36(3)*, 297-308.

Kumar, H., & Kim, H.-G. (2012). Semantically enriched clustered user interest profile built from users' tweets. *Asia Information Retrieval Symposium* (pp. 406-416). Springer, Berlin, Heidelberg.

Kumar, H., & Kim, H.-G. (2012). Semantically enriched user interest profile built from users' tweets. *International Conference on Asian Digital Libraries* (pp. 333-337). Springer, Berlin, Heidelberg.

Laere, S., Buyl, R., & Nyssen, M. (2014). A method for detecting behavior-based user profiles in collaborative ontology engineering. *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 657-673). Springer, Berlin, Heidelberg.

Laere, S., Buyl, R., Nyssen, M., & Debruyne, C. (2017). Detecting User Profiles in Collaborative Ontology Engineering Using a User's Interactions. *Journal on Data Semantics, Vol.6(2)*, 71-82.

Lampos, V., Aletras, N., Geyti, J., Zou, B., & Cox, I. (2016). Inferring the socioeconomic status of social media users based on behaviour and language. *European Conference on Information Retrieval* (pp. 689-695). Springer, Cham.

Lee, J., Hussain, R., Rivera, V., & Isroilov, D. (2018). Second-level degree-based entity resolution in online social networks. *Social Network Analysis and Mining, Vol.8(1)*, 1-8.

Lee, T.-S., & Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems With Applications, Vol.28(4)*, 743-752.

Lenail, A. (2020, 07 07). *NN SVG*. Retrieved from http://alexlenail.me/NN-SVG/index.html

Leong, C. (2016). Credit Risk Scoring with Bayesian Network Models. *Computational Economics, Vol.47(3)*, 423-446.

Li, Y., Yang, L., Xu, B., Wang, J., & Lin, H. (2019). Improving User Attribute Classification with Text and Social Network Attention. *Cognitive Computation, Vol.11(4)*, 459-468.

Li, Z., Guo, B., Sun, Y., Wang, Z., Wang, L., & Yu, Z. (2019). An attention-based user profiling model by leveraging multi-modal social media contents. *Communications in Computer and Information Science, Vol.1138*, 272-284.

Liao, L., Huang, H., & Wang, Y. (2015). Multi-roles affiliation model for general user profiling. *International Conference on Database Systems for Advanced Applications* (pp. 227-233). Springer, Cham.

Liberati, C., & Camillo, F. (2018). Personal values and credit scoring: new insights in the financial prediction. *Journal of the Operational Research Society, Vol.69(12)*, 1994-2005.

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., & Li, X. (2019). Efficient User Profiling Based Intelligent Travel Recommender System for

Individual and Group of Users. *Mobile Networks and Applications, Vol.24(3)*, 1018-1033.

Lu, L. (2016). Private Banks in China: Origin, Challenges and Regulatory Implications. *Banking & Finance Law Review, 31(3), 585.*

Lully, V., Laublet, P., Stankovic, M., & Radulovic, F. (2018). Image user profiling with knowledge graph and computer vision. *European Semantic Web Conference* (pp. 100-104). Springer, Cham.

M, G., V, A., S, M.-M., & H, M. (2020). Concept attribution: Explaining CNN decisions to physicians. *Computers in biology and medicine, Vol.123*.

Ma, C., Zhu, C., Fu, Y., Zhu, H., Liu, G., & Chen, E. (2015). Social user profiling: A social-aware topic modeling perspective. *International Conference on Database Systems for Advanced Applications* (pp. 610-622). Springer, Cham.

Ma, J., Wen, J., Zhong, M., Chen, W., & Li, X. (2019). MMM: Multi-source Multi-net Micro-video Recommendation with Clustered Hidden Item Representation Learning. *Data Science and Engineering, Vol.4(3)*, 240-253.

Manca, M., Boratto, L., & Carta, S. (2018). Behavioral data mining to produce novel and serendipitous friend recommendations in a social bookmarking system. *Information Systems Frontiers, Vol.20(4)*, 825-839.

Mavri, M., Angelis, V., Ioannou, G., Gaki, E., & Koufodontis, I. (2008). A two-stage dynamic credit scoring model, based on customers' profile and time horizon. *Journal of Financial Services Marketing, Vol.13(1)*, 17.

Mayring, P. (2004). Qualitative content analysis. *A companion to qualitative research, 1*, 159-176.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3(2)*.

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of Social Media Research: Common Concerns and Practical Considerations. *Cyberpsychology, behavior, and social networking, 16(9)*, pp. 708-713.

Most famous social network sites worldwide. (2017, 10 04). Retrieved from Statista: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America, Vol.116(44)*, 22071-22080.

Nicoletti, M., Schiaffino, S., & Godoy, D. (2013). Mining interests for user profiling in electronic conversations. *Expert Systems With Applications, Vol.40(2)*, 638-645.

Nijhawan, L. P., Janodia, M. D., Muddukrishna, B. S., Bhat, K. M., Bairy, K. L., Udupa, N., & Musmade, P. B. (2013). Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research, 4(3)*, p. 134.

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction, 27(3-5)*, 393-444.

Orlandi, F. (2012). Multi-source provenance-aware user interest profiling on the social semantic web. *Proceedings of the 20th international conference on user modeling, adaptation, and personalization*, (pp. 378-381).

Our Products. (2018, 01 25). Retrieved from Leveraging Technology Solutions in Credit and Verification | Lenddo: https://www.lenddo.com/products.html

Packin, N. G., & Lev-Aretz, Y. (2016). On Social Credit and the Right to Be Unnetworked. *Columbia Business Law Review*, 339-544.

Pang, G., Jiang, S., & Chen, D. (2013). A simple integration of social relationship and text data for identifying potential customers in microblogging. *International Conference on Advanced Data Mining and Applications* (pp. 397-409). Springer, Berlin, Heidelberg.

Panigutti, C., Perotti, A., & Pedreschi, D. (2020). Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (pp. 629-639).

Peña, P., Del Hoyo, R., Vea-Murguía, J., González, C., & Mayo, S. (2013). Automatic ontology user profiling for social networks from URLs shared. *Conference of the Spanish Association for Artificial Intelligence* (pp. 168-177). Springer, Berlin, Heidelberg.

Peng, J., Detchon, S., Choo, K., & Ashman, H. (2017). Astroturfing detection in social media: a binary n-gram-based approach. *Concurrency And Computation-Practice & Experience, Vol.29(17)*.

Pereira, F., Gama, J., Amo, S., & Oliveira, G. (2018). On analyzing user preference dynamics with temporal social networks. *Machine Learning, Vol.107(11)*, 1745-1773.

Piao, G., & Breslin, J. (2018). Inferring user interests in microblogging social networks: a survey. *User Modeling and User-Adapted Interaction, Vol.28(3)*, 277-329.

Pintelas, E., Liaskos, M., Livieris, I. E., Kotsiantis, S., & Pintelas, P. (2020). Explainable Machine Learning Framework for Image Classification Problems: Case Study on Glioma Cancer Prediction. *Journal of imaging, Vol.6(37)*, 37.

Pintelas, E., Livieris, I. E., & Pintelas, P. (2020). A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability. *Algorithms, Vol.13(1)*, 17.

Pipanmaekaporn, L., & Kamonsantiroj, S. (2015). A belief function reasoning approach to web user profiling. *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 509-516). Springer, Cham.

Pla Karidi, D., Stavrakas, Y., & Vassiliou, Y. (2018). Tweet and followee personalized recommendations based on knowledge graphs. *Journal of Ambient Intelligence and Humanized Computing, Vol.9(6)*, 2035-2049.

Preece, A. (2018). Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management, Vol.25(2)*, 63-72.

Regulation (EU) 2016/679. (2020). Retrieved March 1, 2021, from EUR-Lex: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

Relevant Data Protection Laws. (2020). Retrieved from Austrian Data Protection Authority: https://www.data-protection-authority.gv.at/relevant-data-protection-laws

Reznik, M. (2012). Identity theft on social networking sites: Developing issues of internet impersonation. *Touro Law Review, 29*, p. 455.

Rio-Torto, I., Fernandes, K., & Teixeira, L. F. (2020). Understanding the decisions of CNNs: An in-model approach. *Pattern Recognition Letters, Vol.133*, 373-380.

RIS Informationsangebote. (2020). Retrieved from The Legal Information System of the Republic of Austria: https://www.ris.bka.gv.at/

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access, Vol.8*, 42200-42216.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems, Vol.33(6)*, 673-705.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Prentice Hall.

Search Engine Market Share Worldwide | StatCounter Global Stats. (2020, April). Retrieved from StatCounter Global Stats: https://gs.statcounter.com/search-engine-market-share

Shabtai, A., Bercovitch, M., Rokach, L., & Elovici, Y. (2014). Optimizing data misuse detection. *ACM Transactions on Knowledge Discovery from Data (TKDD), 8(3), 16*.

Shan, W. (2017). Chinese Society in 2016: Stable but under Tightened Control. *East Asian Policy, 9(01)*, pp. 63-77.

Sheh, R., & Monteath, I. (2018). Defining Explainable AI for Requirements Analysis. *KI-Künstliche Intelligenz, 32(4)*, 261-266.

Shen, F., Wang, R., & Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy, Vol.26(2)*.

Shi, J., & Xu, B. (2016). Credit Scoring by Fuzzy Support Vector Machines with a Novel Membership Function. *Journal of Risk and Financial Management, Vol.9(4)*.

Shi, J., Zhang, S.-y., & Qiu, L.-m. (2013). Credit scoring by feature-weighted support vector machines. *Journal of Zhejiang University SCIENCE C, Vol.14(3)*, 197-204.

Siami, M., Gholamian, M. R., & Basiri, J. (2014). An application of locally linear model tree algorithm with combination of feature selection in credit scoring. *International Journal of Systems Science, Vol.45(10)*, 2213-2222.

Sibona, C. (2014). Unfriending on Facebook: Context collapse and unfriending behaviors. *In 2014 47th Hawaii International Conference on System Sciences* (pp. 1676-1685). IEEE.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable Deep Learning Models in Medical Image Analysis. *Journal of imaging, Vol.6(52)*, 52.

Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2020). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics, Vol.26(1)*, 1064-1074.

Sreejesh, S. M. (2014). Business research design: Exploratory, descriptive and causal designs. *Business Research Methods* (pp. 25-103). Springer, Cham.

Stoughton, W. J., Thompson, L. F., & Meade, A. W. (2015). Examining applicant reactions to the use of social networking websites in pre-employment screening. *Journal of Business and Psychology, 30(1)*, 73-88.

Sultana, M., Paul, P., & Gavrilova, M. (2016). Identifying users from online interactions in twitter. *Transactions on Computational Science XXVI* (pp. 111-124). Springer, Berlin, Heidelberg.

Syed Mustapha, S. (2018). Case-based reasoning for identifying knowledge leader within online community. *Expert Systems With Applications, Vol.97*, 244-252.

Ta, N., Li, G.-L., Hu, J., & Feng, J.-H. (2019). Location and Trajectory Identification from Microblogs. *Journal of Computer Science and Technology, Vol.34(4)*, 727-746.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). LINE: Large-scale Information Network Embedding. *Proceedings of the 24th international conference on world wide web*, (pp. 1067-1077).

Tang, J., Yao, L., Zhang, D., & Zhang, J. (2010). A Combination Approach to Web User Profiling. *ACM Transactions on Knowledge Discovery from Data (TKDD), Vol.5(1)*, 1-44.

Tang, X., Xu, Y., & Geva, S. (2014). Tensor Reduction for User Profiling in Personalized Recommender Systems. *Proceedings of the 2014 Australasian Document Computing Symposium*, 34-41.

Tao, Y., & Zhang, W. (2016). Establishment of cross-border e-commerce credit evaluation system based on big data. *Management & Engineering, (24), 3.*

The Austrian Data Protection Authority. (2020). Retrieved from https://www.dsb.gv.at/

Toch, E., Wang, Y., & Cranor, L. F. (2012). Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction, 22(1-2)*, pp. 203-220.

Tomczak, J. M., & Zięba, M. (2015). Classification Restricted Boltzmann Machine for comprehensible credit scoring model. *Expert Systems With Applications, Vol.42(4)*, 1789-1796.

Tong, P., Yao, J., Wang, L., & Yang, S. (2016). Comprehensive graph and content feature based user profiling. *Australasian Database Conference* (pp. 31-42). Springer, Cham.

Trepte, S. (2020). The social media privacy model: Privacy and communication in the light of social media affordances. *Communication Theory*, 1-22.

TU CatalogPlus Search. (2020). Retrieved from TU Wien: https://catalogplus.tuwien.ac.at/

Turculeț, M. (2014). Ethical issues concerning online social networks. *Procedia-Social and Behavioral Sciences 149*, pp. 967-972.

Usman, M. (2015). Supporting Effort Estimation in Agile Software Development. *Doctoral dissertation*. Blekinge Tekniska Högskola.

Usman, M., Britto, R., Börstler, J., & Mendes, E. (2017). Taxonomies in software engineering: A systematic mapping study and a revised taxonomy development method. *Information and Software Technology, 85*, 43-59.

Valsamis, A., Psychas, A., Aisopos, F., Menychtas, A., & Varvarigou, T. (2017). Second screen user profiling and multi-level smart recommendations in the context of social TVs. *International Symposium on Emerging Technologies for Education* (pp. 514-525). Springer, Cham.

Van Selm, M. &. (2006). Conducting online surveys. *Quality and quantity, 40(3)*, 435-456.

Waltl, B., & Vogl, R. (2018). Increasing Transparency in Algorithmic- Decision-Making with Explainable AI. *Datenschutz und Datensicherheit-DuD, 42(10)*, 613-617.

Wang, Q., Ma, S., & Zhang, C. (2017). Predicting users' demographic characteristics in a Chinese social media network. *The Electronic Library, Vol.35(4)*, 758-769.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., & Cranor, L. F. (2011). I regretted the minute I pressed share: A qualitative study of regrets on Facebook. *In Proceedings of the seventh symposium on usable privacy and security* (p. 10). ACM.

Wang, Y., Zhong, Z., Yang, A., & Jing, N. (2018). A deep point-of-interest recommendation system in location-based social networks. *International Conference on Data Mining and Big Data* (pp. 547-554). Springer, Cham.

Won, H.-H., Myung, W., Song, G.-Y., Lee, W.-H., Kim, J.-W., Carroll, B. J., & Kim, D. K. (2013). Predicting national suicide numbers with social media data. *PloS one, 8(4)*.

Wright, K. B. (2006). Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of computer-mediated communication, 10(3), JCMC1034.*

Wu, L., Wang, D., Guo, C., Zhang, J., & Chen, C. (2016). User profiling by combining topic modeling and pointwise mutual information (TM-PMI). *International Conference on Multimedia Modeling* (pp. 152-161). Springer, Cham.

Xie, Q., Wang, Y., Xu, Z., Yu, K., Wei, C., & Yu, Z. (2018). First Place Solution for NLPCC 2018 Shared Task User Profiling and Recommendation. *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 16-26). Springer, Cham.

Xu, B., Tadesse, M., Fei, P., & Lin, H. (2019). Multi-granularity convolutional neural network with feature fusion and refinement for user profiling. *China Conference on Information Retrieval* (pp. 161-172). Springer, Cham.

Xu, D., Cui, P., Zhu, W., & Yang, S. (2014). Graph-Based Residence Location Inference for Social Media Users. *IEEE MultiMedia, Vol.21(4)*, 76-83.

Xu, R., Du, J., Zhao, Z., He, Y., Gao, Q., & Gui, L. (2019). Inferring user profiles in social media by joint modeling of text and networks. *Science China Information Sciences, Vol.62(11)*, 1-3.

Yang, D., Xiao, Y., Tong, H., Zhang, J., & Wang, W. (2015). An integrated tag recommendation algorithm towards Weibo user profiling. *International Conference on Database Systems for Advanced Applications* (pp. 353-373). Springer, Cham.

Yap, B. W., Ong, S. H., & Husain, N. H. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems With Applications, Vol.38(10)*, 13274-13283.

Yin, Y., Thapliya, R., & Zimmermann, R. (2018). Encoded Semantic Tree for Automatic User Profiling Applied to Personalized Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology, Vol.28(1)*, 181-192.

You, Q., Bhatia, S., & Luo, J. (2016). A picture tells a thousand words—About you! User interest profiling from user generated visual content. *Signal Processing, Vol.124*, 45-53.

Zarrinkalam, F., Kahani, M., & Bagheri, E. (2019). User interest prediction over future unobserved topics on social networks. *Information Retrieval Journal, Vol.22(1)*, 93-128.

Zeng, G. (2017). A comparison study of computational methods of Kolmogorov-Smirnov statistic in credit scoring. *Communications in Statistics - Simulation and Computation, Vol.46(10)*, 7744-7760.

Zhang, H., Zeng, R., Chen, L., & Zhang, S. (2020). Research on personal credit scoring model based on multi-source data. *Journal of Physics: Conference Series, Vol.1437*, 012053.

Zhang, L., Fu, S., Jiang, S., Bao, R., & Zeng, Y. (2018). A Fusion Model of Multi-data Sources for User Profiling in Social Media. *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 3-15). Springer, Cham.

Zhang, S., Xiong, W., Ni, W., & Li, X. (2015). Value of big data to finance: observations on an internet credit Service Company in China. *Financial Innovation, 1(1), 17*.

Zhang, Z., & Bors, G. (2019). "Less is more": Mining useful features from Twitter user profiles for Twitter user classification in the public health domain. *Online Information Review, Vol.44(1)*, 213-237.

Zheng, X., Luo, Y., Sun, L., Zhang, J., & Chen, F. (2018). A tourism destination recommender system using users' sentiment and temporal dynamics. *Journal of Intelligent Information Systems, Vol.51(3)*, 557-578.

Zheng, Y., Li, L., Zhang, J., Xie, Q., & Zhong, L. (2019). Using sentiment representation learning to enhance gender classification for user profiling. *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data* (pp. 3-11). Springer, Cham.

Zhou, L., Lai, K., & Yu, L. (2009). Credit scoring using support vector machines with direct search for parameters selection. *Soft Computing, Vol.13(2)*, 149-155.

Zhou, X., Wang, W., & Jin, Q. (2015). Multi-dimensional attributes and measures for dynamical user profiling in social networking environments. *Multimedia Tools and Applications, Vol.74(14)*, 5015-5028.

Zhou, X., Xu, Y., Li, Y., Josang, A., & Cox, C. (2012). The state-of-the-art in personalized recommender systems for social networking. *Artificial Intelligence Review, Vol.37(2)*, 119-132.

Zhuang, C., Ma, Q., & Yoshikawa, M. (2017). SNS user classification and its application to obscure POI discovery. *Multimedia Tools and Applications, Vol.76(4)*, 5461-5487.

Zihni, E., Madai, V. I., Livne, M., Galinovic, I., Khalil, A. A., Fiebach, J. B., & Frey, D. (2020). Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PloS one, Vol.15(4)*, e0231166.

Zuiderveen, B. F., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., . . . de Vreese, C. (2018). Online political microtargeting: promises and threats for democracy. *Utrecht Law Review, 14(1)*, pp. 82-96.

Appendices

Appendix A. Credit scoring model components SLR search results

Title	Author(s)	Year
Classification Restricted Boltzmann Machine for comprehensible credit scoring model	Tomczak, Jakub M; Zięba, Maciej	2015
A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model	Chi, Bo-Wen; Hsu, Chiun-Chieh	2012
A Credit Scoring Model for Microfinance Bank Based on Fuzzy Classifier Optimized by a Differential Evolution Algorithm	Baklouti, Ibtissem	2014
Constructing a reassigning credit scoring model	Chuang, Chun-Ling; Lin, Rong-Ho	2009
A two-stage dynamic credit scoring model, based on customers' profile and time horizon	Maria Mavri; Vassilis Angelis; George Ioannou; Eleni Gaki; Iason Koufodontis	2008
Managing loan customers using misclassification patterns of credit scoring model	Kim, Yoon Seong; Sohn, So Young	2004
Research on personal credit scoring model based on multi-source data	Zhang, Haichao; Zeng, Ruishuang; Chen, Linling; Zhang, Shangfeng	2020
A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach	Feng Shen; Run Wang; Yu Shen	2020
Application of the Scoring Model for Assessing the Credit Rating of Principals	Margarita Janeska; Suzana Taleska; Kosta Sotiroski	2014
A two-stage least cost credit scoring model	Gehrlein, William; Wagner, Bret	1997

A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines	Lee, Tian-Shyug; Chen, I-Fei	2005
Credit scoring by feature-weighted support vector machines	Shi, Jian; Zhang, Shu-you; Qiu, Le-miao	2013
A Comparison of a Credit Scoring Model with a Credit Performance Model	Crook, J.N; Hamilton, R; Thomas, L.C	1992
Hybrid mining approach in the design of credit scoring models	Hsieh, Nan-Chen	2005
An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data	Akkoç, Soner	2012
Credit scoring in mortgage lending: evidence from China	Gan, Christopher; Li, Zhaohua; Wang, Weizhuo; Kao, Betty	2012
Credit Risk Assessment Model Based Using Principal component Analysis And Artificial Neural Network	Hamdy Abeer; Hussein Walid B	2016
A hybrid neural network approach for credit scoring	Chuang, Chun-Ling; Huang, Szu-Teng	2011
Credit Scoring by Fuzzy Support Vector Machines with a Novel Membership Function	Shi, Jian; Xu, Benlian	2016
Using data mining to improve assessment of credit worthiness via credit scoring models	Yap, Bee Wah; Ong, Seng Huat; Husain, Nor Huselina Mohamed	2011
A psychological approach to microfinance credit scoring via a classification and regression tree	Baklouti, Ibtissem	2014
Credit scoring for individuals	Maria Dimitriu; Elena Alexandra Avramescu; Razvan Constantin Caracota	2010
Using Bayesian networks to perform reject inference	Anderson, Billie	2019

Credit scoring using support vector machines with direct search for parameters selection	Zhou, Ligang; Lai, Kin; Yu, Lean	2009
Credit scoring, augmentation and lean models	Banasik, J; Crook, J	2005
Exploring the Nature of Credit Scoring: A Neuro Fuzzy Approach	Akkoç, Soner	2019
The applicability of credit scoring models in emerging economies: an evidence from Jordan	Abbod, Maysam; Radi, Mohammed	2018
The Consumer Loan's Payment Default Predictive Model: an Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank	Abid, Lobna; Masmoudi, Afif; Zouari-Ghorbel, Sonia	2018
The Consumer Loan's Payment Default Predictive Model: An Application in a Tunisian Commercial Bank	Abid, Lobna; Masmoudi, Afif; Zouari-Ghorbel, Sonia	2016
A comparison study of computational methods of Kolmogorov-Smirnov statistic in credit scoring	Zeng, Guoping	2017
Credit Scoring Using Data Mining Techniques	Koh, Hian; Tan, Wei	2004
Would credit scoring work for Islamic finance? A neural network approach	Abdou, Hussein; Alam, Shaair; Mulkeen, James	2014
Credit Risk Scoring with Bayesian Network Models	Leong, Chee	2016
Bayesian Network Modeling: A Case Study of Credit Scoring Analysis of Consumer Loans Default Payment	Abid, Lobna; Zaghdene, Soukeina; Masmoudi, Afif	2017
An evaluation of alternative scoring models in private banking	Abdou, Hussein A	2009
An application of locally linear model tree algorithm with combination of feature selection in credit scoring	Siami, Mohammad; Gholamian, Mohammad Reza; Basiri, Javad	2014
What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance	De Cnudde, Sofie; Moeyersoms, Julie; Stankova, Marija; Tobback, Ellen; Javaly, Vinayak; Martens, David	2019

Credit scoring processes from a knowledge management perspective	Kiss, Ferenc	2003
From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring	Guo, Guangming; Zhu, Feida; Chen, Enhong; Liu, Qi; Wu, Le; Guan, Chu	2016
Personal values and credit scoring: new insights in the financial prediction	Liberati, Caterina; Camillo, Furio	2018

Table 14. Credit scoring model components SLR search results

Appendix B. Credit scoring model components extracted terms

accommodation type (Banasik & Crook, 2005)**, account balance** (Zeng, 2017)**, active early repayment status** (Zhang, Zeng, Chen, & Zhang, 2020)**, additional income** (Abdou H. A., 2009)**, age** (Abbod & Radi, 2018) (Abdou H. A., 2009) (Abdou, Alam, & Mulkeen, 2014) (Abeer & B, 2016) (Abid, Masmoudi, & Zouari-Ghorbel, 2016) (Abid, Masmoudi, & Zouari-Ghorbel, 2018) (Abid, Zaghdene, & Masmoudi, 2017) (Anderson, 2019) (Baklouti, 2014a) (Baklouti, 2014a) (Chi & Hsu, 2012) (Dimitriu, Avramescu, & Caracota, 2010) (Gan, Li, Wang, & Kao, 2012) (Gehrlein & Wagner, 1997) (Janeska, Taleska, & Sotiroski, 2014) (Koh & Tan, 2004) (Lee & Chen, 2005) (Leong, 2016) (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008) (Shen, Wang, & Shen, 2020) (Tomczak & Zięba, 2015) (Zeng, 2017) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Yap, Ong, & Husain, 2011) (Zhou, Lai, & Yu, 2009) (Banasik & Crook, 2005)**, amount and time of deferred repayment of specific business** (Zhang, Zeng, Chen, & Zhang, 2020)**, amount and time of prepayment of specific business** (Zhang, Zeng, Chen, & Zhang, 2020)**, amount of the monthly installment** (Kiss, 2003)**, average consumption and maximum consumption of credit card in the past 6 months** (Zhang, Zeng, Chen, & Zhang, 2020)**, average loan repayment time** (Zhang, Zeng, Chen, & Zhang, 2020)**, balance of current account** (Tomczak & Zięba, 2015), **bank accounts** (Kiss, 2003)**, banking activity** (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008)**, banking references at bank X** (Dimitriu, Avramescu, & Caracota, 2010)**, banking relationship** (Gehrlein & Wagner, 1997)**, basic working type** (Akkoç, 2012)**, better-than-average** (Baklouti, 2014a)**, borrowing interest rate** (Shen, Wang, & Shen, 2020)**, branch** (Abdou H. A., 2009)**, business sector** (Baklouti, 2014a) (Baklouti, 2014a)**, capacity** (Siami, Gholamian, & Basiri, 2014)**, capital** (Siami, Gholamian, & Basiri, 2014)**, car asset** (Shen, Wang, & Shen, 2020)**, Car exists** (Akkoç, 2012)**, car loans** (Shen, Wang, & Shen, 2020)**, character** (Siami, Gholamian, & Basiri, 2014)**, charge card** (Crook, Hamilton, &

Thomas, 1992)**, checking account** (Chuang & Lin, 2009) (Kim & Sohn, 2004) (Abeer & B, 2016)**, cheque card** (Crook, Hamilton, & Thomas, 1992)**, children** (Chi & Hsu, 2012)**, co-applicant information** (Gehrlein & Wagner, 1997)**, collateral** (Siami, Gholamian, & Basiri, 2014)**, company** (Abdou H. A., 2009)**, company size** (Shen, Wang, & Shen, 2020)**, company type** (Shen, Wang, & Shen, 2020)**, concurrent credits** (Zeng, 2017)**, conditions** (Siami, Gholamian, & Basiri, 2014)**, corporate guarantee** (Abdou H. A., 2009)**, credit amount** (Akkoç, 2012) (Zeng, 2017) (Abid, Zaghdene, & Masmoudi, 2017) (Baklouti, 2014a) (Baklouti, 2014a) (Chuang & Lin, 2009) (Chuang & Lin, 2009) (Hsieh, 2005) (Kim & Sohn, 2004) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009) (Abeer & B, 2016)**, credit card history overdue total days** (Zhang, Zeng, Chen, & Zhang, 2020)**, credit card overdue amount** (Zhang, Zeng, Chen, & Zhang, 2020)**, credit card repayment speed** (Zhang, Zeng, Chen, & Zhang, 2020)**, credit card status** (Abdou H. A., 2009)**, credit certification** (Shen, Wang, & Shen, 2020)**, credit duration** (Abid, Zaghdene, & Masmoudi, 2017) (Chuang & Lin, 2009), **credit history** (Akkoç, 2012) (Abeer & B, 2016) (Chuang & Lin, 2009) (Chuang & Lin, 2009) (Hsieh, 2005) (Janeska, Taleska, & Sotiroski, 2014) (Kim & Sohn, 2004) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**, credit purpose** (Baklouti, 2014a) (Baklouti, 2014a)**, credit references** (Gehrlein & Wagner, 1997)**, credit status of guarantor** (Lee & Chen, 2005)**, current account** (Crook, Hamilton, & Thomas, 1992)**, current electoral roll category** (Banasik & Crook, 2005)**, debt ratio** (Gehrlein & Wagner, 1997)**, debtor or guarantor of credit granted by another institution** (Abeer & B, 2016)**, debt-to-income ratio** (Chi & Hsu, 2012)**, defaulters/non-defaulters** (Yap, Ong, & Husain, 2011)**, degree of indebtedness for the applicant family** (Dimitriu, Avramescu, & Caracota, 2010)**, delinquency status in the last 3-6-12 months** (Chi & Hsu, 2012)**, deposit account** (Crook, Hamilton, & Thomas, 1992)**, district of address** (Yap, Ong, & Husain, 2011)**, duration** (Gan, Li, Wang, & Kao, 2012)**, duration in current address in years** (Zeng, 2017)**, Duration in month** (Akkoç, 2012) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009) (Kim & Sohn, 2004) (Zeng, 2017) (Abeer & B, 2016)**, education** (Baklouti, 2014a) (Chi & Hsu, 2012) (Gan, Li, Wang, & Kao, 2012) (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008) (Abdou H. A., 2009) (Akkoç, 2012) (Shen, Wang, & Shen, 2020) (Baklouti, 2014a) (Lee & Chen, 2005)**, emotional intelligence** (Baklouti, 2014a)**, employment status** (Crook, Hamilton, & Thomas, 1992)**, estimated value of home** (Crook, Hamilton, & Thomas, 1992)**, European credit card** (Anderson, 2019)**, financial credibility** (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008)**, financial statement** (Kiss, 2003)**, floor space of the property** (Chi & Hsu, 2012)**, forced early repayment status** (Zhang, Zeng, Chen, & Zhang, 2020)**, foreign worker** (Akkoç, 2012) (Zeng, 2017) (Abeer & B, 2016) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**,**

gender (Abbod & Radi, 2018) (Abdou H. A., 2009) (Abdou, Alam, & Mulkeen, 2014) (Abid, Zaghdene, & Masmoudi, 2017) (Akkoç, 2012) (Baklouti, 2014a) (Baklouti, 2014a) (Chi & Hsu, 2012) (Koh & Tan, 2004) (Lee & Chen, 2005) (Leong, 2016) (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008) (Shen, Wang, & Shen, 2020) (Yap, Ong, & Husain, 2011)**, guarantors** (Zeng, 2017)**, guarantors or collateral** (Kiss, 2003)**, guarantors/other debtors** (Chuang & Lin, 2009)**, holder of other credit cards** (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008)**, home duration** (Akkoç, 2012)**, home telephone** (Abdou H. A., 2009) (Baklouti, 2014a)**, house owned or rented** (Abdou H. A., 2009)**, house rent > loan tenure** (Abdou H. A., 2009)**, housing** (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**, housing arrangements** (Abeer & B, 2016)**, housing status** (Abdou, Alam, & Mulkeen, 2014) (Gehrlein & Wagner, 1997) (Janeska, Taleska, & Sotiroski, 2014)**, identity certification** (Shen, Wang, & Shen, 2020)**, illusion of control** (Baklouti, 2014a)**, income** (Anderson, 2019) (Chi & Hsu, 2012) (Crook, Hamilton, & Thomas, 1992) (Gan, Li, Wang, & Kao, 2012) (Gehrlein & Wagner, 1997) (Leong, 2016) (Koh & Tan, 2004)**, income certification** (Shen, Wang, & Shen, 2020)**, income statement** (Kiss, 2003)**, industry** (Shen, Wang, & Shen, 2020)**, installment rate in percentage of disposable income** (Akkoç, 2012) (Chuang & Lin, 2009) (Hsieh, 2005)**, instalment per cent** (Zeng, 2017)**, interest-based look-a-likes** (De Cnudde, et al., 2019)**, job** (Abeer & B, 2016) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**, job/position** (Kiss, 2003)**, job designation** (Leong, 2016)**, job experience** (Baklouti, 2014a) (Baklouti, 2014a)**, job status** (Chi & Hsu, 2012)**, length of existing relationship** (Gan, Li, Wang, & Kao, 2012)**, length of residence** (Leong, 2016)**, length of service** (Janeska, Taleska, & Sotiroski, 2014)**, obligations to his bank** (Abbod & Radi, 2018)**, life insurance policies** (Kiss, 2003)**, live in urban or rural area** (Anderson, 2019)**, living area** (Chi & Hsu, 2012)**, loan amount** (Abbod & Radi, 2018) (Abid, Masmoudi, & Zouari-Ghorbel, 2016) (Abdou H. A., 2009) (Abdou, Alam, & Mulkeen, 2014) (Abid, Masmoudi, & Zouari-Ghorbel, 2018) (Chi & Hsu, 2012) (Lee & Chen, 2005) (Shen, Wang, & Shen, 2020) (Anderson, 2019)**, loan amount/house appraisal value** (Lee & Chen, 2005)**, loan and credit card account types and the number of credit accounts of each type** (Zhang, Zeng, Chen, & Zhang, 2020)**, loan class** (Abbod & Radi, 2018)**, loan duration** (Abbod & Radi, 2018) (Abdou H. A., 2009) (Abdou, Alam, & Mulkeen, 2014)**, loan period** (Dimitriu, Avramescu, & Caracota, 2010) (Shen, Wang, & Shen, 2020)**, loan purpose** (Abbod & Radi, 2018) (Zhou, Lai, & Yu, 2009)**, loan repayment data from banks** (Kiss, 2003)**, loan status for accepted applicants** (Anderson, 2019)**, loan type** (Lee & Chen, 2005)**, loans from other banks** (Abbod & Radi, 2018) (Abdou H. A., 2009)**, loan-to-value ratio** (Chi & Hsu, 2012) (Gan, Li, Wang, & Kao, 2012)**, locative situation** (Dimitriu, Avramescu, & Caracota, 2010)**, major credit card** (Crook, Hamilton, & Thomas, 1992)**, marital status** (Abbod & Radi, 2018) (Abdou H. A., 2009) (Abdou, Alam, & Mulkeen, 2014) (Akkoç, 2012) (Baklouti, 2014a) (Baklouti, 2014a)

(Dimitriu, Avramescu, & Caracota, 2010) (Gan, Li, Wang, & Kao, 2012) (Koh & Tan, 2004) (Lee & Chen, 2005) (Leong, 2016) (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008) (Yap, Ong, & Husain, 2011) (Chi & Hsu, 2012)**, maturity of the mortgage** (Chi & Hsu, 2012)**, miscalibration** (Baklouti, 2014a)**, mode of income** (Abdou, Alam, & Mulkeen, 2014)**, mode of work** (Abdou, Alam, & Mulkeen, 2014)**, monthly expense** (Abdou, Alam, & Mulkeen, 2014)**, monthly income** (Abdou, Alam, & Mulkeen, 2014) (Lee & Chen, 2005) (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008) (Shen, Wang, & Shen, 2020)**, monthly installment/monthly income** (Lee & Chen, 2005)**, monthly net income** (Janeska, Taleska, & Sotiroski, 2014) (Dimitriu, Avramescu, & Caracota, 2010)**, monthly repayment burden** (Abid, Zaghdene, & Masmoudi, 2017)**, monthly salary** (Abbod & Radi, 2018) (Abdou H. A., 2009)**, mortgage** (Shen, Wang, & Shen, 2020)**, mortgage balance outstanding** (Crook, Hamilton, & Thomas, 1992)**, most valuable available asset** (Zeng, 2017)**, number in household** (Anderson, 2019)**, number of cars** (Yap, Ong, & Husain, 2011)**, number of children** (Anderson, 2019) (Crook, Hamilton, & Thomas, 1992) (Koh & Tan, 2004) (Leong, 2016)**, number of children under 16** (Banasik & Crook, 2005)**, number of credit cards held** (Zhang, Zeng, Chen, & Zhang, 2020)**, number of dependents** (Abdou, Alam, & Mulkeen, 2014) (Dimitriu, Avramescu, & Caracota, 2010) (Gehrlein & Wagner, 1997) (Yap, Ong, & Husain, 2011) (Zeng, 2017) (Crook, Hamilton, & Thomas, 1992)**, number of existing credits at this bank** (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009) (Abeer & B, 2016) (Zeng, 2017)**, number of finished loans** (Anderson, 2019)**, number of guarantors** (Lee & Chen, 2005) (Chi & Hsu, 2012)**, number of institutions** (Zhang, Zeng, Chen, & Zhang, 2020)**, number of loans** (Zhang, Zeng, Chen, & Zhang, 2020)**, number of other credit cards held** (Koh & Tan, 2004)**, number of outstanding loans** (Anderson, 2019)**, number of outstanding loans at bank** (Anderson, 2019)**, number of people being liable to provide maintenance for** (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009) (Abeer & B, 2016)**, number of previous loans** (Baklouti, 2014a) (Baklouti, 2014a)**, number of times and the overdue duration** (Zhang, Zeng, Chen, & Zhang, 2020)**, number of searches in last 6 months** (Banasik & Crook, 2005)**, number of years spent at present residence** (Abeer & B, 2016)**, number of years spent at the current address** (Kiss, 2003) (Kiss, 2003)**, occupation** (Zeng, 2017) (Abdou, Alam, & Mulkeen, 2014) (Chi & Hsu, 2012) (Gan, Li, Wang, & Kao, 2012) (Lee & Chen, 2005) (Yap, Ong, & Husain, 2011)**, occupation code** (Banasik & Crook, 2005)**, occupation group** (Gehrlein & Wagner, 1997)**, occupational category** (Abid, Masmoudi, & Zouari-Ghorbel, 2016) (Abid, Masmoudi, & Zouari-Ghorbel, 2018)**, other debtors / guarantors** (Akkoç, 2012) (Zhou, Lai, & Yu, 2009) (Hsieh, 2005) (Kim & Sohn, 2004) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Chuang & Lin, 2009)**, other guarantors** (Abdou H. A., 2009)**, other installment plans** (Akkoç, 2012) (Abeer & B, 2016) (Chuang & Lin,

2009) (Kim & Sohn, 2004) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009) (Hsieh, 2005)**, outgoings** (Crook, Hamilton, & Thomas, 1992)**, outstanding bills** (Leong, 2016)**, outstanding credit** (Abid, Masmoudi, & Zouari-Ghorbel, 2016) (Abid, Masmoudi, & Zouari-Ghorbel, 2018) (Abid, Zaghdene, & Masmoudi, 2017)**, overdraft** (Abdou, Alam, & Mulkeen, 2014)**, overdue repayment** (Zhang, Zeng, Chen, & Zhang, 2020)**, own contribution** (Dimitriu, Avramescu, & Caracota, 2010)**, own property** (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008)**, owns a car** (Abdou, Alam, & Mulkeen, 2014) (Leong, 2016)**, owns a house or not** (Abbod & Radi, 2018)**, payment status of previous credit** (Zeng, 2017)**, period of time in the same work** (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008)**, person status and sex** (Zhou, Lai, & Yu, 2009)**, personal status and sex** (Abeer & B, 2016) (Akkoç, 2012) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013)**, phone** (Crook, Hamilton, & Thomas, 1992)**, phone at residence** (Gehrlein & Wagner, 1997)**, position held** (Kiss, 2003)**, preferred credibility limit** (Mavri, Angelis, Ioannou, Gaki, & Koufodontis, 2008)**, present employment** (Kim & Sohn, 2004) (Akkoç, 2012) (Chuang & Lin, 2009) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**, present employment status** (Abeer & B, 2016)**, present residence** (Kim & Sohn, 2004) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**, previous credits** (Tomczak & Zięba, 2015), **previous loans default** (Baklouti, 2014a) (Baklouti, 2014a)**, profession** (Abid, Zaghdene, & Masmoudi, 2017) (Anderson, 2019)**, property** (Chuang & Lin, 2009) (Hsieh, 2005) (Kim & Sohn, 2004) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**, property age** (Lee & Chen, 2005)**, property type** (Lee & Chen, 2005)**, proximity to bank** (Gan, Li, Wang, & Kao, 2012)**, proximity towards bank X branches** (Dimitriu, Avramescu, & Caracota, 2010)**, purpose** (Akkoç, 2012) (Zeng, 2017) (Chuang & Lin, 2009) (Chuang & Lin, 2009) (Hsieh, 2005) (Kim & Sohn, 2004) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Abeer & B, 2016) (Lee & Chen, 2005) (Chi & Hsu, 2012) (Tomczak & Zięba, 2015), **race** (Yap, Ong, & Husain, 2011)**, real estate** (Shen, Wang, & Shen, 2020)**, region** (Akkoç, 2012)**, region of country where applicant lives** (Anderson, 2019)**, region of living** (Gan, Li, Wang, & Kao, 2012)**, relation with other banks** (Abdou H. A., 2009)**, relation with the bank** (Janeska, Taleska, & Sotiroski, 2014)**, relational look-a-likes** (De Cnudde, et al., 2019)**, relationship between guarantor and guarantee** (Lee & Chen, 2005)**, repayment type** (Shen, Wang, & Shen, 2020)**, residence type** (Akkoç, 2012) (Anderson, 2019)**, residential status** (Crook, Hamilton, & Thomas, 1992)**, saving account/bonds** (Zhou, Lai, & Yu, 2009)**, savings account** (Chuang & Lin, 2009) (Kim & Sohn, 2004)**, savings account/bonds** (Akkoç, 2012) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support

vector machines, 2013) (Abeer & B, 2016) (Chuang & Lin, 2009)**, semiometric space variables** (Liberati & Camillo, 2018)**, sex & marital status** (Zeng, 2017)**, sex** (Kiss, 2003)**, social network** (De Cnudde, et al., 2019)**, socio-demographic data** (De Cnudde, et al., 2019)**, special loan for government employees** (Lee & Chen, 2005)**, spouse's income** (Crook, Hamilton, & Thomas, 1992)**, status of existing checking account** (Akkoç, 2012) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Zhou, Lai, & Yu, 2009) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013)**, telephone** (Banasik & Crook, 2005) (Zeng, 2017) (Abeer & B, 2016) (Chuang & Lin, 2009) (Hsieh, 2005) (Shi & Xu, 2016) (Shi, Zhang, & Qiu, Credit scoring by feature-weighted support vector machines, 2013) (Zhou, Lai, & Yu, 2009)**, television area code** (Banasik & Crook, 2005)**, time at address** (Anderson, 2019)**, time at current/previous job** (Gehrlein & Wagner, 1997)**, time at current/previous residence** (Gehrlein & Wagner, 1997)**, time at present address** (Banasik & Crook, 2005)**, time from the first loan and first credit card to the current time** (Zhang, Zeng, Chen, & Zhang, 2020)**, time of inquiry by the organization and corresponding reasons** (Zhang, Zeng, Chen, & Zhang, 2020)**, time on job** (Anderson, 2019)**, title/position** (Abdou H. A., 2009)**, total current work duration** (Akkoç, 2012)**, total historical loan amount** (Zhang, Zeng, Chen, & Zhang, 2020)**, total work duration** (Akkoç, 2012)**, type of apartment** (Zeng, 2017)**, type of bank/building society accounts** (Banasik & Crook, 2005)**, type of business** (Anderson, 2019)**, type of credit** (Abdou, Alam, & Mulkeen, 2014) (Abid, Zaghdene, & Masmoudi, 2017)**, type of credit cards** (Anderson, 2019)**, type of job** (Abbod & Radi, 2018)**, type of property possessed by client** (Abeer & B, 2016)**, type of residence** (Leong, 2016)**, type of transportation** (Anderson, 2019)**, use of the loan** (Shen, Wang, & Shen, 2020)**, User Demographic Attributes** (Guo, et al., 2016)**, User-Generated Content** (Guo, et al., 2016)**, User Social Network** (Guo, et al., 2016)**, utility bill** (Abdou H. A., 2009)**, value of property** (Chi & Hsu, 2012)**, value savings/stocks** (Zeng, 2017)**, weeks since last county court judgement** (Banasik & Crook, 2005)**, whether an applicant has an outstanding mortgage loan** (Koh & Tan, 2004)**, width of current employment in years** (Zeng, 2017)**, work certification** (Shen, Wang, & Shen, 2020)**, work sector** (Yap, Ong, & Husain, 2011)**, work seniority** (Chi & Hsu, 2012)**, working period within the last employer** (Dimitriu, Avramescu, & Caracota, 2010)**, working time** (Shen, Wang, & Shen, 2020)**, years at bank** (Crook, Hamilton, & Thomas, 1992)**, years at present job** (Abbod & Radi, 2018)**, years in present employment** (Crook, Hamilton, & Thomas, 1992)**, years on electoral roll at current address** (Banasik & Crook, 2005)**, years working at the current company** (Lee & Chen, 2005)**.**

Appendix C. Credit scoring model components terminology control

active early repayment status, age, amount and time of deferred repayment of specific business, amount and time of repayment of specific business, amount of the monthly instalment, average consumption and maximum consumption of credit card in the past 6 month, average credit repayment time, balance of the current account (account balance, balance of current account)**, balance on checking account(s)** (checking account, status of existing checking account)**, balance on savings (deposit) account(s)** (deposit account, saving account/bonds, savings account, savings account/bonds, value savings/stocks)**, bank accounts, banking activity, better-than-average, borrowing interest rate, business sector (industry)** (business sector, industry, type of business, work sector)**, capacity, capital, car asset** (car asset, car exists, number of cars, owns a car, type of transportation)**, character, charge card, cheque card, co-applicant information, collateral, company, company size, company type, credit amount** (credit amount, loan amount)**, credit card history overdue total days, credit card overdue amount, credit card repayment speed, credit card status** (credit card status, European credit card, holder of other credit cards, major credit card)**, credit certificate, credit conditions** (conditions, instalment per cent, instalment rate in percentage of disposable income, loan class, monthly instalment/monthly income)**, credit duration** (credit duration, duration, duration in month, loan duration, loan period)**, credit history** (credit history, loan repayment data from banks, previous credits, payment status of previous credit)**, credit type** (loan type, loan status for accepted applicants, special loan for government employees, type of credit)**, credit purpose** (credit purpose, loan purpose, purpose, use of the loan)**, credit status of guarantor, credit repayment type** (conditions, repayment type)**, credit-to-value ratio** (house rent > loan tenure, loan amount/house appraisal value, loan-to-value ratio)**, current car credit** (car loans)**, current electoral roll category, current income** (additional income, income, income certification, income statement, monthly income, monthly net income, monthly salary)**, current mortgage** (maturity of the mortgage, mortgage, mortgage balance outstanding, whether an applicant has an outstanding mortgage loan)**, debt ratio, debt-to-income ratio, defaulter/non-defaulter (past defaults)** (defaulters/non-defaulters)**, delinquency status in the last 3-6-12 months, education, emotional intelligence, ethnicity** (race)**, expenses** (outgoings, outstanding bills, utility bill)**, financial credibility** (financial credibility, financial statement), **floor space of the property, forced early repayment status, gender** (gender, person status and sex, personal status and sex, sex & marital status, sex)**, housing** (accommodation type, housing, housing arrangements, housing status, present residence, residence type, residential status, type of residence, type of apartment)**, home duration** (home duration, duration in current address in years)**, identity certification, illusion of control, interest-based look-a-likes, job** (job, present employment)**, job experience** (job experience, length of service, period of time in the same work, time at

current/previous job, total current work duration, time on job, years at present job, years in present employment, years working at the current company, width of current employment in years, working period within the last employer, working time)**, job status** (job status, employment status, present employment status, work certification)**, job title/position** (job, job/position, job designation, position held, title/position)**, length of relationship (years at bank)** (length of existing relationship)**, life insurance policies, marital status** (marital status, person status and sex, personal status and sex, sex & marital status, socio-demographic data)**, miscalibration, mode of income, mode of work, monthly expenses, monthly repayment burden, most valuable available asset, number in household, number of children (under 16)** (children, number of children, number of children under 16)**, number of current credits** (concurrent credits, loans from other banks, number of loans)**, number of dependents** (number of dependents, number of people being liable to provide maintenance for)**, number of existing credits at this bank, number of guarantors** (credit references, number of guarantors)**, number of institutions with present credits from** (number of institutions)**, number of previous credits** (number of finished loans, number of previous loans)**, number of searches in last 6 months, number of times and overdue duration, occupation** (occupation, profession)**, occupation group** (basic working type, occupation code, occupation group, occupational category, type of job)**, other debtors (guarantors)** (guarantors, guarantors or collateral, guarantors/other debtors, other debtors / guarantors, other guarantors, debtor or guarantor of credit granted by another institution, degree of indebtedness for the applicant family, corporate guarantee)**, other instalment plans (credits)** (loan and credit card account types and the number of credit accounts of each type, number of outstanding loans, number of outstanding loans at bank, obligations to his bank, other instalment plans)**, outstanding credit, overdraft, overdue repayment, own contribution (as per credit purpose)** (own contribution)**, preferred credibility limit (credit cards)** (preferred credibility limit)**, credit cards account types** (loan and credit card account types and the number of credit accounts of each type, number of credit cards held, number of other credit cards held)**, present residence** (district of address, foreign worker)**, previous credits defaults, proximity to bank** (proximity to bank, proximity towards bank X branches)**, real estate** (own property, owns a house or not, property, property age, property type, real estate, type of property possessed by client, house owned or rented)**, relation with bank** (banking references at bank X, banking relationship, branch, relation with the bank, relationship between guarantor and guarantee, type of bank/building society accounts, years at bank)**, relation with other banks, relational look-a-likes, residence region** (district of address, live in urban or rural area, living area, locative situation, region, region of country where applicant lives, region of living)**, semiometric space variables, social network, socio-demographic data** (socio-demographic data, User Demographic Attributes)**, spouse's income, telephone** (home telephone, phone, phone at residence, telephone)**, television area code, time at current/previous residence** (length of residence, number of years spent at present residence, number of years spent at the current

address, time at address, time at current/previous residence, time at present address)**, time at previous job** (time at current/previous job)**, time from the first credit card to the current time (i.e., credit time limit class)** (time from the first loan and first credit card to the current time)**, time from the first credit to the current time (i.e., credit time limit class)** (time from the first loan and first credit card to the current time)**, time of inquiry by the organization and corresponding reasons, total historical credit amount, total working duration, type of account(s)** (current account)**, type of credit card(s)** (type of credit cards)**, user-generated content, user social network, value of property** (value of property, estimated value of home)**, weeks since last county court judgement, work seniority, years on electoral roll at current address**.

Appendix D. Credit scoring model components

active early repayment status, age, amount and time of deferred repayment of specific business, amount and time of repayment of specific business, amount of the monthly instalment, average consumption and maximum consumption of credit card in the past 6 months, average credit repayment time, balance of the current account, balance on checking account(s), balance on savings (deposit) account(s), bank accounts, banking activity, better-than-average, borrowing interest rate, business sector (industry), capacity, capital, car asset, character, charge card, cheque card, co-applicant information, collateral, company, company size, company type, credit amount, credit card history overdue total days, credit card overdue amount, credit card repayment speed, credit card status, credit certificate, credit conditions, credit duration, credit history, credit type, credit purpose, credit status of guarantor, credit repayment type, credit-to-value ratio, current car credit, current electoral roll category, current income, current mortgage, debt ratio, debt-to-income ratio, defaulter/non-defaulter (past defaults), delinquency status in the last 3-6-12 months, education, emotional intelligence, ethnicity, expenses, financial credibility, floor space of the property, forced early repayment status, gender, housing, home duration, identity certification, illusion of control, interest-based look-a-likes, job, job experience, job status, job title/position, length of relationship (years at bank), life insurance policies, marital status, miscalibration, mode of income, mode of work, monthly expenses, monthly repayment burden, most valuable available asset, number in household, number of children (under 16), number of current credits, number of dependents, number of existing credits at this bank, number of guarantors, number of institutions with present credits from, number of previous credits, number of searches in last 6 months, number of times and overdue duration, occupation, occupation group, other debtors (guarantors), other instalment plans (credits), outstanding credit, overdraft, overdue repayment, own contribution (as per credit purpose), preferred credibility limit (credit cards), credit cards account types, present residence, previous credits

defaults, proximity to bank, real estate, relation with bank, relation with other banks, relational look-a-likes, residence region, semiometric space variables, social network, socio-demographic data, spouse's income, telephone, television area code, time at current/previous residence, time at previous job, time from the first credit card to the current time (i.e., credit time limit class), time from the first credit to the current time (i.e., credit time limit class), time of inquiry by the organization and corresponding reasons, total historical credit amount, total working duration, type of account(s), type of credit card(s), user-generated content, user social network, value of property, weeks since last county court judgement, work seniority, years on electoral roll at current address.

Appendix E. Social media user profiling SLR search results

Title	Author(s)	Year
Social user profiling: A social-aware topic modeling perspective	Ma, C.; Zhu, C.; Fu, Y.; Zhu, H.; Liu, G.; Chen, E.	2015
An integrated tag recommendation algorithm towards Weibo user profiling	Yang, D.; Xiao, Y.; Tong, H.; Zhang, J.; Wang, W.	2015
Mining interests for user profiling in electronic conversations	Nicoletti, Matias; Schiaffino, Silvia; Godoy, Daniela	2013
Image user profiling with knowledge graph and computer vision	Lully, V.; Laublet, P.; Stankovic, M.; Radulovic, F.	2018
Comprehensive graph and content feature based user profiling	Tong, P.; Yao, J.; Wang, L.; Yang, S.	2016
Accurate online social network user profiling	Dougnon, R.Y.; Fournier-Viger, P.; Lin, J.C.-W.; Nkambou, R.	2015
Multi-roles affiliation model for general user profiling	Liao, L.; Huang, H.; Wang, Y.	2015
Folksonomy-based fuzzy user profiling for improved recommendations	Anand, Deepa; Mampilli, Bonson Sebastian	2014
User profiling via affinity-aware friendship network	Chen, Z.; Zhu, F.; Guo, G.; Liu, H.	2014

User profiling based on keyword clusters for improved recommendations	Anand, D.; Mampilli, B.S.	2014
Automatic ontology user profiling for social networks from URLs shared	Peña, P.; Del Hoyo, R.; Vea-Murguía, J.; González, C.; Mayo, S.	2013
A Combination Approach to Web User Profiling	Tang, Jie; Yao, Limin; Zhang, Duo; Zhang, Jing	2010
Forum User Profiling by Incorporating User Behavior and Social Network Connections	Chen, D.; Zhang, Q.; Chen, G.; Fan, C.; Gao, Q.	2018
A belief function reasoning approach to web user profiling	Pipanmaekaporn, L.; Kamonsantiroj, S.	2015
First Place Solution for NLPCC 2018 Shared Task User Profiling and Recommendation	Xie, Q.; Wang, Y.; Xu, Z.; Yu, K.; Wei, C.; Yu, Z.C.	2018
A Fusion Model of Multi-data Sources for User Profiling in Social Media	Zhang, L.; Fu, S.; Jiang, S.; Bao, R.; Zeng, Y.	2018
Discerning individual interests and shared interests for social user profiling	Chen, Enhong; Zeng, Guangxiang; Luo, Ping; Zhu, Hengshu; Tian, Jilei; Xiong, Hui	2017
User profiling by combining topic modeling and pointwise mutual information (TM-PMI)	Wu, L.; Wang, D.; Guo, C.; Zhang, J.; Chen, C.W.	2016
Twitter user profiling model based on temporal analysis of hashtags and social interactions	Gorrab, A.; Kboubi, F.; Jaffal, A.; Le Grand, B.; Ghezala, H.B.	2017
NLPCC 2018 Shared Task User Profiling and Recommendation Method Summary by DUTIR_9148	Chen, X.; Wang, J.; Ren, Y.; Liu, T.; Lin, H.	2018
Second screen user profiling and multi-level smart recommendations in the context of social TVs	Valsamis, A.; Psychas, A.; Aisopos, F.; Menychtas, A.; Varvarigou, T.	2017
Multi-dimensional attributes and measures for dynamical user profiling in social networking environments	Zhou, Xiaokang; Wang, Wei; Jin, Qun	2015

Encoded Semantic Tree for Automatic User Profiling Applied to Personalized Video Summarization	Yin, Yifang; Thapliya, Roshan; Zimmermann, Roger	2018
User profiling of the Twitter Social Network during the impeachment of Brazilian President	França, Fabrício; Goya, Denise; Camargo Penteado, Claudio	2018
User profiling for big social media data using standing ovation model	Al-Qurishi, Muhammad; Alhuzami, Saad; AlRubaian, Majed; Hossain, M. Shamim; Alamri, Atif; Rahman, Md.	2018
Tensor Reduction for User Profiling in Personalized Recommender Systems	Tang, Xiaoyu; Xu, Yue; Geva, Shlomo	2014
Using sentiment representation learning to enhance gender classification for user profiling	Zheng, Y.; Li, L.; Zhang, J.; Xie, Q.; Zhong, L.	2019
Multi-granularity convolutional neural network with feature fusion and refinement for user profiling	Xu, B.; Tadesse, M.M.; Fei, P.; Lin, H.	2019
A method for detecting behavior-based user profiles in collaborative ontology engineering	Laere, S.V.; Buyl, R.; Nyssen, M.	2014
A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions	Eke, Christopher Ifeanyi; Norman, Azah Anir; Shuib, Liyana; Nweke, Henry Friday	2019
Efficient User Profiling Based Intelligent Travel Recommender System for Individual and Group of Users	Logesh, R.; Subramaniyaswamy, V.; Vijayakumar, V.; Li, Xiong	2019
Combining behaviors and demographics to segment online audiences: Experiments with a youtube channel	Jansen, B.J.; Jung, S.-G.; Salminen, J.; An, J.; Kwak, H.	2018
Multi-view personality profiling based on longitudinal data	Buraya, K.; Farseev, A.; Filchenkov, A.	2018
Semantically enriched user interest profile built from users' tweets	Kumar, H.; Kim, H.-G.	2012

Semantically enriched clustered user interest profile built from users' tweets	Kumar, H.; Kim, H.-G.	2012
Identifying users from online interactions in twitter	Sultana, M.; Paul, P.P.; Gavrilova, M.	2016
A deep approach for multi-modal user attribute modeling	Huang, X.; Yang, Z.; Yang, Y.; Shen, F.; Xie, N.; Shen, H.T.	2017
Inferring the socioeconomic status of social media users based on behaviour and language	Lampos, V.; Aletras, N.; Geyti, J.K.; Zou, B.; Cox, I.J.	2016
A deep point-of-interest recommendation system in location-based social networks	Wang, Y.; Zhong, Z.; Yang, A.; Jing, N.	2018
SNS user classification and its application to obscure POI discovery	Zhuang, Chenyi; Ma, Qiang; Yoshikawa, Masatoshi	2017
Profiling Web users using big data	Gu, Xiaotao; Yang, Hong; Tang, Jie; Zhang, Jing; Zhang, Fanjin; Liu, Debing; Hall, Wendy; Fu, Xiao	2018
Discovery of spatio-temporal patterns from location-based social networks	Béjar, J; Álvarez, S; García, D; Gómez, I; Oliva, L; Tejeda, A; Vázquez-Salceda, J	2016
A picture tells a thousand words—About you! User interest profiling from user generated visual content	You, Quanzeng; Bhatia, Sumit; Luo, Jiebo	2016
Astroturfing detection in social media: a binary n-gram-based approach	Peng, J; Detchon, S; Choo, Kkr; Ashman, H	2017
From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring	Guo, Guangming; Zhu, Feida; Chen, Enhong; Liu, Qi; Wu, Le; Guan, Chu	2016
A simple integration of social relationship and text data for identifying potential customers in microblogging	Pang, G.; Jiang, S.; Chen, D.	2013
Building profiles of blog users based on comment graph analysis: The Habrahabr.ru case	Barysheva, A.; Petrov, M.; Yavorskiy, R.	2015

Detecting User Profiles in Collaborative Ontology Engineering Using a User's Interactions	Laere, Sven; Buyl, Ronald; Nyssen, Marc; Debruyne, Christophe	2017
Discovering Homophily in Online Social Networks	De Salve, Andrea; Guidi, Barbara; Ricci, Laura; Mori, Paolo	2018
An attention-based user profiling model by leveraging multi-modal social media contents	Li, Z.; Guo, B.; Sun, Y.; Wang, Z.; Wang, L.; Yu, Z.	2019
Youtube user and usage profiling: Stories of political horror and security success	Kandias, M.; Mitrou, L.; Stavrou, V.; Gritzalis, D.	2014
Graph-Based Residence Location Inference for Social Media Users	Dan Xu; Peng Cui; Wenwu Zhu; Shiqiang Yang	2014
Second-level degree-based entity resolution in online social networks	Lee, JooYoung; Hussain, Rasheed; Rivera, Victor; Isroilov, Davlatbek	2018
Intelligent travel information platform based on location base services to predict user travel behavior from user-generated GPS traces	Arain, Qasim Ali; Memon, Hina; Memon, Imran; Memon, Muhammad Hammad; Shaikh, Riaz Ahmed; Mangi, Farman Ali	2017
Inferring user interests in microblogging social networks: a survey	Piao, Guangyuan; Breslin, John	2018
Twitter user modeling based on indirect explicit relationships for personalized recommendations	Alshammari, A.; Kapetanakis, S.; Polatidis, N.; Evans, R.; Alshammari, G.	2019
Inferring social network user profiles using a partial social graph	Dougnon, Raïssa; Fournier-Viger, Philippe; Lin, Jerry; Nkambou, Roger	2016
Relational User Attribute Inference in Social Media	Quan Fang; Jitao Sang; Changsheng Xu; Hossain, M. Shamim	2015
The state-of-the-art in personalized recommender systems for social networking	Zhou, Xujuan; Xu, Yue; Li, Yuefeng; Josang, Audun; Cox, Clive	2012
Case-based reasoning for identifying knowledge leader within online community	Syed Mustapha, S.M.F.D	2018

Modeling Topics and Behavior of Microbloggers: An Integrated Approach	Hoang, TA; Lim, Ep	2017
Multi-source provenance-aware user interest profiling on the social semantic web	Orlandi, Fabrizio	2012
Improving User Attribute Classification with Text and Social Network Attention	Li, Yumeng; Yang, Liang; Xu, Bo; Wang, Jian; Lin, Hongfei	2019
Ontology-based user profile learning	Eyharabide, Victoria; Amandi, Analía	2012
On the quality of semantic interest profiles for onine social network consumers	Besel, Christoph; Schlötterer, Jörg; Granitzer, Michael	2016
A social recommender system using deep architecture and network embedding	C C, Nisha; Mohan, Anuraj	2019
MMM: Multi-source Multi-net Micro-video Recommendation with Clustered Hidden Item Representation Learning	Ma, Jingwei; Wen, Jiahui; Zhong, Mingyang; Chen, Weitong; Li, Xue	2019
Deep recurrent convolutional networks for inferring user interests from social media	Kang, Jaeyong; Choi, HongSeok; Lee, Hyunju	2019
User interest prediction over future unobserved topics on social networks	Zarrinkalam, Fattane; Kahani, Mohsen; Bagheri, Ebrahim	2019
Authorship verification applied to detection of compromised accounts on online social networks	Barbon, Sylvio; Igawa, Rodrigo; Bogaz Zarpelão, Bruno	2017
A novel relationship strength model for online social networks	Ju, Chunhua; Tao, Wanqiong	2017
A tourism destination recommender system using users' sentiment and temporal dynamics	Zheng, Xiaoyao; Luo, Yonglong; Sun, Liping; Zhang, Ji; Chen, Fulong	2018
Social recommendations for personalized fitness assistance	Dharia, Saumil; Eirinaki, Magdalini; Jain, Vijesh; Patel, Jvalant; Varlamis, Iraklis; Vora, Jainikkumar; Yamauchi, Rizen	2018

Cognitive computing for customer profiling: meta classification for gender prediction	Hirt, Robin; Kühl, Niklas; Satzger, Gerhard	2019
Location and Trajectory Identification from Microblogs	Ta, Na; Li, Guo-Liang; Hu, Jun; Feng, Jian-Hua	2019
Determining the interests of social media users: two approaches	Bennacer Seghouani, Nacéra; Jipmo, Coriane; Quercini, Gianluca	2019
On analyzing user preference dynamics with temporal social networks	Pereira, Fabíola; Gama, João; Amo, Sandra; Oliveira, Gina	2018
Location prediction in large-scale social networks: an in-depth benchmarking study	Al Hasan Haldar, Nur; Li, Jianxin; Reynolds, Mark; Sellis, Timos; Yu, Jeffrey	2019
Inferring user profiles in social media by joint modeling of text and networks	Xu, Ruifeng; Du, Jiachen; Zhao, Zhishan; He, Yulan; Gao, Qinghong; Gui, Lin	2019
Recommendation of microblog users based on hierarchical interest profiles	Faralli, Stefano; Stilo, Giovanni; Velardi, Paola	2015
Tweet and followee personalized recommendations based on knowledge graphs	Pla Karidi, Danae; Stavrakas, Yannis; Vassiliou, Yannis	2018
“Less is more”: Mining useful features from Twitter user profiles for Twitter user classification in the public health domain	Zhang, Ziqi; Bors, Georgica	2019
Behavioral data mining to produce novel and serendipitous friend recommendations in a social bookmarking system	Manca, Matteo; Boratto, Ludovico; Carta, Salvatore	2018
A multi-source integration framework for user occupation inference in social media systems	Huang, Yanxiang; Yu, Lele; Wang, Xiang; Cui, Bin	2015
Predicting users’ demographic characteristics in a Chinese social media network	Wang, Qiangbing; Ma, Shutian; Zhang, Chengzhi	2017

Table 15. Social media user profiling SLR search results

Appendix F. Social media user profiling approaches extracted terms

affiliation graph model (network structure, probabilistic graphical model) (Liao, Huang, & Wang, 2015), **affinity propagation clustering** (Béjar, et al., 2016)**, analysis of variance (ANOVA)** (Laere, Buyl, & Nyssen, 2014)**, applied standardization (clustering)** (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, association rule mining** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, association use mining** (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, asynchronous stochastic gradient algorithm (ASGD)** (Xu, et al., 2019)**, attention mechanism** (Xu, et al., 2019)**, author topic model** (Huang, et al., 2017)**, autoencoder-based social recommender system (AESR)** (C C & Mohan, 2019)**, autoencoders and perceptrons** (Ma, Wen, Zhong, Chen, & Li, 2019)**, averaging and stacking models** (Tong, Yao, Wang, & Yang, 2016)**, bag of concepts** (Zarrinkalam, Kahani, & Bagheri, 2019)**, bag-of-words (BOW)** (Zhang, Fu, Jiang, Bao, & Zeng, 2018)**, bag-of-words or topic modeling** (Zarrinkalam, Kahani, & Bagheri, 2019)**, based on standing ovation model (SOM)** (Al-Qurishi, et al., 2018)**, Bayesian classification** (Zhang & Bors, 2019)**, Bayesian inference** (Lee, Hussain, Rivera, & Isroilov, 2018)**, Bayesian network** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, Bayesian networks (belief networks)** (C C & Mohan, 2019)**, Bayesian networks and ontologies** (Eyharabide & Amandi, 2012)**, Bayesian personalized ranking** (C C & Mohan, 2019) (Xie, et al., 2018), **Bayesian technique** (Arain, et al., 2017)**, belief function reasoning** (Pipanmaekaporn & Kamonsantiroj, 2015)**, bi-directional gated recurrent unit (biGRU, type of RNN)** (Kang, Choi, & Lee, 2019)**, Bidirectional LSTM (BiLSTM)** (Zheng, Li, Zhang, Xie, & Zhong, 2019) (Zhang & Bors, 2019)**, binary relevance (BR) (user tags prediction)** (Xie, et al., 2018)**, bird flocking** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, Business Semantics Management** (Laere, Buyl, & Nyssen, 2014)**, CALGARI and KL-divergence scored** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019)**, CART tree based model** (Xu, Tadesse, Fei, & Lin, 2019)**, CBR (case-based reasoning)** (Syed Mustapha, 2018)**, CENE (network embedding and content)** (Chen, Zhang, Chen, Fan, & Gao, 2018)**, centroid-based classification (text classification)** (Pang, Jiang, & Chen, 2013)**, CF (collaborative filtering)** (Yang, Xiao, Tong, Zhang, & Wang, 2015)**, Chinese restaurant process (statistics/probability)** (Xu, Tadesse, Fei, & Lin, 2019)**, classic rank (CLR)** (Arain, et al., 2017)**, classification** (França, Goya, & Camargo Penteado, 2018) (Tang, Yao, Zhang, & Zhang, 2010) (Ma, et al., 2015)**, classifier chains (CC)** (Xie, et al., 2018)**, clustering** (Al-Qurishi, et al., 2018) (Anand & Mampilli, 2014) (Béjar, et al., 2016) (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017) (Kumar & Kim, Semantically enriched user interest profile built from users' tweets, 2012) (Kumar & Kim, Semantically enriched clustered user interest profile built from users' tweets, 2012)**, Cosine Similarity** (De Salve, Guidi, Ricci, & Mori, 2018)**, CNN** (Huang, et al., 2017) (Kang, Choi, & Lee, 2019) (Li, Yang, Xu, Wang, & Lin, 2019) (Ma, Wen, Zhong, Chen, & Li, 2019) (Xu, et al., 2019) (Zhang,

Fu, Jiang, Bao, & Zeng, 2018) (Zhang & Bors, 2019) (Zheng, Li, Zhang, Xie, & Zhong, 2019) (You, Bhatia, & Luo, 2016)**, CNN ResNet-50** (Li, et al., 2019)**, CNN to classify images (map to KG)** (Lully, Laublet, Stankovic, & Radulovic, 2018)**, CNN-RNN** (Xie, et al., 2018)**, co-factorization machines (CoFM)** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, collaborative filtering (activity recommendation engine)** (Dharia, et al., 2018)**, collaborative filtering** (Anand & Mampilli, 2014) (Anand & Mampilli, 2014) (C C & Mohan, 2019) (Pla Karidi, Stavrakas, & Vassiliou, 2018) (Tang, Xu, & Geva, 2014) (Wang, Zhong, Yang, & Jing, 2018) (Zarrinkalam, Kahani, & Bagheri, 2019) (Zheng, Luo, Sun, Zhang, & Chen, 2018) (Zhou, Xu, Li, Josang, & Cox, 2012) (Faralli, Stilo, & Velardi, 2015) (Ma, Wen, Zhong, Chen, & Li, 2019)**, Collaborative Filtering (memory-based CF, matrix factorization (SVD, LDA, ALS))** (Xie, et al., 2018)**, collaborative filtering (model-based and memory-based)** (Logesh, Subramaniyaswamy, Vijayakumar, & Li, 2019)**, collaborative filtering rank (CFR)** (Arain, et al., 2017)**, collaborative ontology engineering methods** (Laere, Buyl, & Nyssen, 2014)**, collective classification** (Huang, Yu, Wang, & Cui, 2015)**, frequent pattern mining (FPM) (topic detection)** (Piao & Breslin, 2018)**, collective naïve Bayes** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015)**, combine semantic context and social network information** (Li, Yang, Xu, Wang, & Lin, 2019)**, combined perceptron with Bayes model** (Xu, Tadesse, Fei, & Lin, 2019) (Li, Yang, Xu, Wang, & Lin, 2019)**, community detection** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015) (Liao, Huang, & Wang, 2015)**, composite Gaussian Process (GP)** (Lampos, Aletras, Geyti, Zou, & Cox, 2016)**, compositional recurrent neural network** (Buraya, Farseev, & Filchenkov, 2018)**, condensed filter tree (CFT)** (Xie, et al., 2018)**, constrained label propagation** (You, Bhatia, & Luo, 2016)**, content-based and graph-based features** (Al-Qurishi, et al., 2018)**, content-based and preference-based filtering** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, content-based collaborative filtering** (Anand & Mampilli, 2014)**, content-based collaborative filtering** (Valsamis, Psychas, Aisopos, Menychtas, & Varvarigou, 2017)**, content-based filtering** (C C & Mohan, 2019) (Faralli, Stilo, & Velardi, 2015) (Ma, Wen, Zhong, Chen, & Li, 2019) (Pla Karidi, Stavrakas, & Vassiliou, 2018) (Zhou, Xu, Li, Josang, & Cox, 2012)**, content-based systems** (Anand & Mampilli, 2014) (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, content-based user tag recommendation** (Chen, Wang, Ren, Liu, & Lin, 2018)**, context-aware recommender systems (CARS)** (Tang, Xu, & Geva, 2014)**, continuous bag-of-words (CBOW)** (Kang, Choi, & Lee, 2019)**, co-profiling algorithm** (Chen, Zhu, Guo, & Liu, 2014)**, corr-LDA** (Huang, et al., 2017)**, Cross-Media-LDA (CMLDA)** (Huang, et al., 2017)**, Crisp User Profile based Recommendations (CUP)** (Anand & Mampilli, 2014)**, DBpedia** (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013), **DBSCAN** (Arain, et al., 2017) (Wang, Zhong, Yang, & Jing, 2018)**, decision tree** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016) (Guo, et al., From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, 2016) (Peng, Detchon, Choo, & Ashman, 2017) (Zhang & Bors, 2019)**, decision trees (J48, ADTree,**

REPTree) (Barbon, Igawa, & Bogaz Zarpelão, 2017)**, deep learning** (C C & Mohan, 2019) (Chen, Wang, Ren, Liu, & Lin, 2018) (Huang, et al., 2017)**, deep neural network** (Wang, Zhong, Yang, & Jing, 2018)**, deep neural networks (bi-GRU layer, hierarchical attention layer, BiRNN, concatenation layer)** (Zhang, Fu, Jiang, Bao, & Zeng, 2018)**, DeepWalk** (Chen, Zhang, Chen, Fan, & Gao, 2018) (Tong, Yao, Wang, & Yang, 2016)**, demographic systems** (C C & Mohan, 2019)**, dependence distributions** (Hoang & Lim, 2017)**, DILIGENT** (Laere, Buyl, & Nyssen, 2014)**, dimensionality reduction** (Tong, Yao, Wang, & Yang, 2016)**, dimensionality reduction through network embedding paradigm (matrix factorization)** (C C & Mohan, 2019)**, discriminative influence model** (Chen, Zhu, Guo, & Liu, 2014)**, Dublin Core** (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, Dublin Core Ontology** (Laere, Buyl, & Nyssen, 2014)**, dynamic user attribute model (DUAM)** (Huang, et al., 2017)**, dynamic user clustering topic model (UCT)** (Huang, et al., 2017)**, dynamic weighted ensemble (DWE)** (Zhuang, Ma, & Yoshikawa, 2017)**, EIUCF** (Logesh, Subramaniyaswamy, Vijayakumar, & Li, 2019)**, ensemble learning** (Chen, Zhang, Chen, Fan, & Gao, 2018) (Hirt, Kühl, & Satzger, 2019)**, entropy-based model (EBM)** (Ju & Tao, 2017)**, expert systems** (Syed Mustapha, 2018)**, explicit semantic analysis** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, extend deep autoencoder with top-k semantic social information** (C C & Mohan, 2019)**, factor graph model** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019)**, FCM** (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017)**, Feature Refinement Layer** (Xu, Tadesse, Fei, & Lin, 2019)**, filtering techniques** (Eke, Norman, Shuib, & Nweke, 2019)**, fitness buddies recommendation engine** (Dharia, et al., 2018)**, FOAF** (Laere, Buyl, & Nyssen, 2014) (Laere S. , Buyl, Nyssen, & Debruyne, 2017) (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013) (Tang, Yao, Zhang, & Zhang, 2010) (Xie, et al., 2018)**, formal concept analysis (FCA)** (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017)**, FREQ (frequency of tags)** (Yang, Xiao, Tong, Zhang, & Wang, 2015)**, frequent terms (bag of words)** (Nicoletti, Schiaffino, & Godoy, 2013)**, friends-based collaborative filtering** (Wang, Zhong, Yang, & Jing, 2018)**, fuzzy C-means algorithm for clustering** (Logesh, Subramaniyaswamy, Vijayakumar, & Li, 2019)**, fuzzy logic (OWA (ordered weighted averaging) operators)** (Anand & Mampilli, 2014)**, gated recurrent neural network (CNN-GRU)** (Li, Yang, Xu, Wang, & Lin, 2019)**, Gaussian distribution** (Ma, Wen, Zhong, Chen, & Li, 2019)**, Gaussian Mixture Model** (Xu, Cui, Zhu, & Yang, 2014) (Zhuang, Ma, & Yoshikawa, 2017) (Béjar, et al., 2016)**, Gaussian relational topic model** (Huang, et al., 2017)**, generalist recommender system kernel (GRSK)** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, generalized matrix factorization** (Ma, Wen, Zhong, Chen, & Li, 2019)**, generative influence models** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019)**, generative relationship influence models** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019)**, geographic topic models** (Zhuang, Ma, & Yoshikawa, 2017)**, geographical topic models by utilizing statistical topic models** (Ta, Li, Hu, & Feng, 2019)**, Gibbs sampling in location-based topic models** (Ta, Li, Hu, & Feng, 2019)**, GOSPL** (Laere, Buyl, & Nyssen, 2014) (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, GOSPL with D2RQ**

(Laere, Buyl, & Nyssen, 2014)**, gradient boosted decision trees** (Chen, Zhu, Guo, & Liu, 2014) (Guo, et al., From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, 2016) (Faralli, Stilo, & Velardi, 2015) (Pang, Jiang, & Chen, 2013) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, gradient boosting** (Buraya, Farseev, & Filchenkov, 2018)**, graph based** (Sultana, Paul, & Gavrilova, 2016)**, graph based approaches (centrality, betweenness)** (Kumar & Kim, Semantically enriched user interest profile built from users' tweets, 2012) (Kumar & Kim, Semantically enriched clustered user interest profile built from users' tweets, 2012)**, graph embedding** (Tong, Yao, Wang, & Yang, 2016)**, graph embedding algorithms (LINE, PUHE)** (Chen, Zhang, Chen, Fan, & Gao, 2018)**, graph embedding learning** (Xu, et al., 2019)**, graph partitioning** (Tong, Yao, Wang, & Yang, 2016)**, Graph Theoretic Analysis** (Kandias, Mitrou, Stavrou, & Gritzalis, 2014)**, graph theory** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, graph-based** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019) (Barbon, Igawa, & Bogaz Zarpelão, 2017) (Lee, Hussain, Rivera, & Isroilov, 2018) (Pereira, Gama, Amo, & Oliveira, 2018)**, graph-based (session-based temporal graph)** (Zarrinkalam, Kahani, & Bagheri, 2019)**, graph-based algorithms** (Barysheva, Petrov, & Yavorskiy, 2015) (Chen, Zhu, Guo, & Liu, 2014)**, graph-based user tag recommendation** (Chen, Wang, Ren, Liu, & Lin, 2018)**, graphical models** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, GRU structure** (Li, et al., 2019)**, HCOME** (Laere, Buyl, & Nyssen, 2014)**, heterogenous graph embeddings** (Xu, et al., 2019)**, heuristic approaches (TF, TF-IDF, TI-TextRank, FTF)** (Piao & Breslin, 2018)**, Hidden Markov Models (HMM)** (Jansen, Jung, Salminen, An, & Kwak, 2018)**, hierarchical attention network** (Li, Yang, Xu, Wang, & Lin, 2019) (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, Hierarchical Attention Transfer Network (HATN)** (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, hierarchical Bayesian model** (Chen, Zhang, Chen, Fan, & Gao, 2018) (Wu, Wang, Guo, Zhang, & Chen, 2016)**, hierarchical clustering** (Anand & Mampilli, 2014) (Kumar & Kim, Semantically enriched user interest profile built from users' tweets, 2012) (Kumar & Kim, Semantically enriched clustered user interest profile built from users' tweets, 2012) (Yin, Thapliya, & Zimmermann, 2018)**, hierarchical convolution neural network (CNN)** (Li, Yang, Xu, Wang, & Lin, 2019)**, hierarchical interest graph (from Wikipedia category graph)** (Besel, Schlötterer, & Granitzer, 2016)**, HMM** (Béjar, et al., 2016)**, HMRF (Hidden Markov Random Field)** (Tang, Yao, Zhang, & Zhang, 2010)**, HMRF-KMEANS** (Gu, et al., 2018)**, hypertext induced topic search (HITS)** (Arain, et al., 2017)**, IBM Watson personality insights** (Li, et al., 2019)**, ImageNet/GoogleNet** (Buraya, Farseev, & Filchenkov, 2018)**, incremental Bayesian online updates** (Fang, Sang, Xu, & Hossain, 2015)**, individual filtering (user preferences)** (Dharia, et al., 2018)**, information filtering (IF)** (Pipanmaekaporn & Kamonsantiroj, 2015)**, injected preferences fusion (IPF)** (Zarrinkalam, Kahani, & Bagheri, 2019)**, item-based CF** (Dharia, et al., 2018) (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, ITT** (Anand & Mampilli, 2014)**, k-means** (Alshammari, Kapetanakis, Polatidis, Evans, & Alshammari, 2019) (Arain, et al., 2017) (Barysheva, Petrov, & Yavorskiy, 2015) (Béjar, et al.,

2016) (C C & Mohan, 2019) (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017) (Hoang & Lim, 2017) (Laere S. , Buyl, Nyssen, & Debruyne, 2017) (Ma, Wen, Zhong, Chen, & Li, 2019)**, k-means clustering** (Ju & Tao, 2017) (Laere, Buyl, & Nyssen, 2014)**, k-NN** (Arain, et al., 2017) (Barbon, Igawa, & Bogaz Zarpelão, 2017) (C C & Mohan, 2019) (Li, Yang, Xu, Wang, & Lin, 2019) (Pang, Jiang, & Chen, 2013) (Peng, Detchon, Choo, & Ashman, 2017) (Valsamis, Psychas, Aisopos, Menychtas, & Varvarigou, 2017)**, knowledge graphs** (Piao & Breslin, 2018) (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, knowledge-based systems (KB)** (C C & Mohan, 2019)**, label propagation** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015) (Li, Yang, Xu, Wang, & Lin, 2019) (You, Bhatia, & Luo, 2016) (Xu, Tadesse, Fei, & Lin, 2019)**, labeled-LDA** (Hoang & Lim, 2017) (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, language models** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019)**, large-scale information network embedding (LINE)** (Chen, Wang, Ren, Liu, & Lin, 2018)**, latent factor model** (Chen, et al., 2017) (Tang, Xu, & Geva, 2014) (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, Latent Semantic Analysis** (Li, et al., 2019) (Wang, Ma, & Zhang, 2017)**, Latent Semantic Analysis (LSA) using matrix factorization technique** (Kumar & Kim, Semantically enriched user interest profile built from users' tweets, 2012) (Kumar & Kim, Semantically enriched clustered user interest profile built from users' tweets, 2012)**, latent semantic hashing** (C C & Mohan, 2019)**, latent SVM (LSVM)** (Fang, Sang, Xu, & Hossain, 2015)**, LDA** (Al-Qurishi, et al., 2018) (Béjar, et al., 2016) (Bennacer Seghouani, Jipmo, & Quercini, 2019) (Buraya, Farseev, & Filchenkov, 2018) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016) (Huang, et al., 2017) (Hoang & Lim, 2017) (Huang, Yu, Wang, & Cui, 2015) (Ju & Tao, 2017) (Kang, Choi, & Lee, 2019) (Pang, Jiang, & Chen, 2013) (Pereira, Gama, Amo, & Oliveira, 2018) (Pla Karidi, Stavrakas, & Vassiliou, 2018) (Wang, Ma, & Zhang, 2017) (Wu, Wang, Guo, Zhang, & Chen, 2016) (Zarrinkalam, Kahani, & Bagheri, 2019) (Zhang & Bors, 2019) (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, Levenshtein Similarity** (De Salve, Guidi, Ricci, & Mori, 2018)**, LINE (large information network embedding)** (Xu, et al., 2019)**, linear kernel SVM** (Zhang & Bors, 2019)**, linear regression (adapted balance winnow algorithm)** (Li, Yang, Xu, Wang, & Lin, 2019)**, linear regression** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, linear SVM** (Fang, Sang, Xu, & Hossain, 2015)**, L-LDA** (Ma, et al., 2015)**, logistic regression** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019) (Buraya, Farseev, & Filchenkov, 2018) (Chen, Zhu, Guo, & Liu, 2014) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016) (Gu, et al., 2018) (Guo, et al., From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, 2016) (Kandias, Mitrou, Stavrou, & Gritzalis, 2014) (Li, Yang, Xu, Wang, & Lin, 2019) (Xu, Tadesse, Fei, & Lin, 2019) (Zhang, Fu, Jiang, Bao, & Zeng, 2018) (Zhang & Bors, 2019) (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, Linguist Quantifier driven Tag Determination (LQT)** (Anand & Mampilli, 2014)**, LSTM** (Buraya, Farseev, & Filchenkov, 2018) (Li, et al., 2019) (Xu, et al., 2019) (Zhang, Fu, Jiang, Bao,

& Zeng, 2018)**, LSTM (sentiment classifier)** (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, LTPA (local tag propagation)** (Yang, Xiao, Tong, Zhang, & Wang, 2015)**, Lucene Clustering** (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017)**, majority vote** (Liao, Huang, & Wang, 2015), **majority voting** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015)**, map image to knowledge graph entities** (Lully, Laublet, Stankovic, & Radulovic, 2018)**, markov chains** (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, Markov logic network (MLN)** (Gu, et al., 2018) (Lee, Hussain, Rivera, & Isroilov, 2018)**, Markov random field** (Chen, Zhu, Guo, & Liu, 2014)**, matrix decomposition techniques (specifically non-negative matrix factorization (NMF))** (Jansen, Jung, Salminen, An, & Kwak, 2018)**, matrix factorization** (Arain, et al., 2017) (C C & Mohan, 2019) (Chen, et al., 2017) (Dharia, et al., 2018) (Zhang, Fu, Jiang, Bao, & Zeng, 2018) (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, Maximum Likelihood Estimation** (Zhuang, Ma, & Yoshikawa, 2017)**, MCMC** (Chen, Zhu, Guo, & Liu, 2014)**, memory-based CF** (Chen, Wang, Ren, Liu, & Lin, 2018)**, ML models stacking (ensemble)** (Xu, Tadesse, Fei, & Lin, 2019)**, MLP (multi-layer perceptron)** (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, mm-LDA** (Huang, et al., 2017)**, modded SVD (modSVD)** (Kumar & Kim, Semantically enriched user interest profile built from users' tweets, 2012) (Kumar & Kim, Semantically enriched clustered user interest profile built from users' tweets, 2012)**, model-based CF** (Chen, Wang, Ren, Liu, & Lin, 2018)**, model-based recommendation (matrix factorization, probabilistic latent factor models)** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, modified balanced winnow algorithm (learning a linear classifier)** (Xu, Tadesse, Fei, & Lin, 2019)**, most popular friends (MPF)** (Xie, et al., 2018)**, MRF (Markov Random Field)** (Tang, Yao, Zhang, & Zhang, 2010)**, multi-granularity CNN** (Xu, Tadesse, Fei, & Lin, 2019)**, multi-modal deep belief network (DBN)** (Huang, et al., 2017)**, multi-modal deep Boltzmann Machines (DBM)** (Huang, et al., 2017)**, multi-model user attribute model (mmUAM)** (Huang, et al., 2017)**, multinominal logistic regression (MLR)** (Kandias, Mitrou, Stavrou, & Gritzalis, 2014)**, Naïve Bayes** (Barbon, Igawa, & Bogaz Zarpelão, 2017) (Buraya, Farseev, & Filchenkov, 2018) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015) (Lee, Hussain, Rivera, & Isroilov, 2018) (Li, et al., 2019) (Ma, et al., 2015) (Pang, Jiang, & Chen, 2013) (Zhang & Bors, 2019)**, naïve Bayes classifier** (Zhou, Xu, Li, Josang, & Cox, 2012)**, naïve Bayes for classification** (Guo, et al., From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, 2016)**, naïve Bayes multinominal (NBM)** (Kandias, Mitrou, Stavrou, & Gritzalis, 2014)**, nearest neighbor distribution over ODP (Open Directory Project)** (Piao & Breslin, 2018)**, neighborhood-based CF** (Tang, Xu, & Geva, 2014)**, neighborhood-based methods** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, network analysis (graph-based)** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, network embeddings** (Zhang, Fu, Jiang, Bao, & Zeng, 2018)**, network representation learning (NRL)** (C C & Mohan, 2019)**, neural network model (social convolution attention neural network)** (Li, Yang, Xu, Wang, & Lin, 2019)**, neural networks** (Li, et al., 2019) (Pla Karidi, Stavrakas, & Vassiliou, 2018) (Xu, et al., 2019)**, neural**

recommendations (neural networks and collaborative filtering) (Ma, Wen, Zhong, Chen, & Li, 2019)**, n-grams** (Chen, Zhang, Chen, Fan, & Gao, 2018)**, N-grams authorship verification (AV)** (Barbon, Igawa, & Bogaz Zarpelão, 2017)**, NLP** (Anand & Mampilli, 2014) (Barysheva, Petrov, & Yavorskiy, 2015) (Hirt, Kühl, & Satzger, 2019) (Kang, Choi, & Lee, 2019) (Xu, Cui, Zhu, & Yang, 2014) (Zhang & Bors, 2019)**, non-linear (radial basis function) kernel SVM** (Zhang & Bors, 2019)**, non-negative matrix factorization (NMF)** (Buraya, Farseev, & Filchenkov, 2018) (C C & Mohan, 2019)**, normalized graphs** (Peng, Detchon, Choo, & Ashman, 2017)**, OKM** (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017)**, ontologies** (Eyharabide & Amandi, 2012) (Piao & Breslin, 2018) (Tang, Yao, Zhang, & Zhang, 2010)**, ontology based** (Eke, Norman, Shuib, & Nweke, 2019)**, ontology engineering project** (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, ontology to categorize results** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, ontology-based recommendations** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, ontology-based user models** (Orlandi, 2012)**, OpenDNS,** (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013)**, outliers determined by interquartile ranges (IQR)** (Peng, Detchon, Choo, & Ashman, 2017)**, OWL** (Laere S. , Buyl, Nyssen, & Debruyne, 2017) (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013)**, PCA (clustering)** (C C & Mohan, 2019) (Laere, Buyl, & Nyssen, 2014) (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, PCFA (principal component factor analysis) (clustering)** (Anand & Mampilli, 2014)**, PDS** (Pipanmaekaporn & Kamonsantiroj, 2015)**, IPE** (Pipanmaekaporn & Kamonsantiroj, 2015)**, PGBN (Poisson Gamma Belief Network, a deep learning topic model)** (Huang, et al., 2017)**, popularity rank (PR)** (Arain, et al., 2017)**, POS (part of speech)** (Chen, Zhang, Chen, Fan, & Gao, 2018)**, probabilistic approaches (Explicit semantic analysis (ESA))** (Piao & Breslin, 2018)**, probabilistic framework** (Chen, Zhu, Guo, & Liu, 2014) (Ta, Li, Hu, & Feng, 2019) (Zarrinkalam, Kahani, & Bagheri, 2019)**, probabilistic inference (for user location)** (Xu, Cui, Zhu, & Yang, 2014)**, probabilistic latent semantic analysis (PLSA)** (Pla Karidi, Stavrakas, & Vassiliou, 2018) (Wu, Wang, Guo, Zhang, & Chen, 2016)**, probabilistic matrix factorization (PMF)** (C C & Mohan, 2019) (Chen, et al., 2017) (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, probabilistic model** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019) (Xie, et al., 2018)**, probabilistic topic model** (Bennacer Seghouani, Jipmo, & Quercini, 2019) (Chen, Zhang, Chen, Fan, & Gao, 2018) (Tang, Yao, Zhang, & Zhang, 2010) (Zhou, Wang, & Jin, 2015)**, probability distributions** (Bennacer Seghouani, Jipmo, & Quercini, 2019) (Piao & Breslin, 2018)**, probability matrix factorization (PMF)** (Wang, Zhong, Yang, & Jing, 2018)**, probability models** (Li, Yang, Xu, Wang, & Lin, 2019)**, Pattern Taxonomy Model (PTM)** (Pipanmaekaporn & Kamonsantiroj, 2015)**, QBLDA** (Hoang & Lim, 2017)**, QICE** (Logesh, Subramaniyaswamy, Vijayakumar, & Li, 2019)**, random forest** (Barbon, Igawa, & Bogaz Zarpelão, 2017) (Gu, et al., 2018) (Zhang & Bors, 2019) (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, random forest classification** (Xu, Tadesse, Fei, & Lin, 2019)**, rating-based systems** (C C & Mohan, 2019)**, Rank based Degree of Feature (RDF)** (Anand & Mampilli, 2014), **RDF** (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013)**,**

recommender systems (social recommendation) (Wu, Wang, Guo, Zhang, & Chen, 2016)**, relational graph** (Lee, Hussain, Rivera, & Isroilov, 2018)**, relational naïve Bayes classifier** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2015) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, replicated softmax model** (Huang, et al., 2017)**, representation learning (feature learning)** (Tong, Yao, Wang, & Yang, 2016)**, Restricted Boltzmann Machines (RBMs)** (Huang, et al., 2017)**, RNN** (Ma, Wen, Zhong, Chen, & Li, 2019) (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, RNN based collaborative filtering** (C C & Mohan, 2019)**, Rocchio** (Pang, Jiang, & Chen, 2013) (Pipanmaekaporn & Kamonsantiroj, 2015)**, rule-based** (Lee, Hussain, Rivera, & Isroilov, 2018)**, rule-based systems** (Gu, et al., 2018)**, SALSA (stochastic approach for link-structure analysis)** (Manca, Boratto, & Carta, 2018)**, semantic methods to recommend friends** (Xie, et al., 2018)**, semantic relationships** (Syed Mustapha, 2018)**, semantic structures** (Al-Qurishi, et al., 2018)**, semantic technologies for interlinking social websites** (Orlandi, 2012)**, semantic trees** (Yin, Thapliya, & Zimmermann, 2018)**, semantically enrich user profiles by using association rules** (Eyharabide & Amandi, 2012)**, semi-supervised topic model** (Ma, et al., 2015)**, sentiment analysis (topic models)** (França, Goya, & Camargo Penteado, 2018)**, similarity-/neighborhood-based** (C C & Mohan, 2019)**, singular value decomposition (SVD)** (C C & Mohan, 2019) (Chen, et al., 2017) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, SIOC** (Laere, Buyl, & Nyssen, 2014) (Laere S. , Buyl, Nyssen, & Debruyne, 2017)**, Smith-Waterman Similarity** (De Salve, Guidi, Ricci, & Mori, 2018)**, social graph** (Yin, Thapliya, & Zimmermann, 2018)**, social pertinent walker (SPTW)** (Logesh, Subramaniyaswamy, Vijayakumar, & Li, 2019) (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, social tagging system (STS)** (Tang, Xu, & Geva, 2014)**, social-based collaborative filtering (CF)** (Chen, Wang, Ren, Liu, & Lin, 2018)**, social-based filtering (friends' preferences)** (Dharia, et al., 2018)**, socially embedded visual representation learning (SEVIR)** (Huang, et al., 2017)**, Softmax layer** (Li, et al., 2019) (Xu, Tadesse, Fei, & Lin, 2019)**, SoRec (social regularization)** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, spectral clustering** (Béjar, et al., 2016) (Lampos, Aletras, Geyti, Zou, & Cox, 2016)**, spreading activation algorithm (e.g., algorithm over semantic networks)** (Besel, Schlötterer, & Granitzer, 2016)**, stacked model to do classifier stacking** (Huang, Yu, Wang, & Cui, 2015)**, stacked SVM** (Fang, Sang, Xu, & Hossain, 2015)**, stacking (ensemble)** (Zhang, Fu, Jiang, Bao, & Zeng, 2018)**, stacking and boosting enhanced ensemble** (Guo, et al., From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, 2016)**, statistical analysis** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, statistical classifier** (Huang, Yu, Wang, & Cui, 2015)**, statistical modeling** (Eke, Norman, Shuib, & Nweke, 2019)**, stochastic gradient descent classifier** (Zhang & Bors, 2019)**, stochastic topic model** (Manca, Boratto, & Carta, 2018)**, supervised topic modeling** (Hoang & Lim, 2017)**, SVD** (Kumar & Kim, Semantically enriched user interest profile built from users' tweets, 2012) (Kumar & Kim, Semantically enriched clustered user interest profile built from users' tweets, 2012) (Wang, Zhong, Yang, & Jing, 2018) (Xu, Tadesse, Fei, & Lin, 2019) (Zhang, Fu, Jiang, Bao, & Zeng,

2018) (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, SVM** (Chen, Zhu, Guo, & Liu, 2014) (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016) (Faralli, Stilo, & Velardi, 2015) (Gu, et al., 2018) (Guo, et al., From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring, 2016) (Kandias, Mitrou, Stavrou, & Gritzalis, 2014) (Lee, Hussain, Rivera, & Isroilov, 2018) (Li, et al., 2019) (Ma, et al., 2015) (Pang, Jiang, & Chen, 2013) (Peng, Detchon, Choo, & Ashman, 2017) (Pipanmaekaporn & Kamonsantiroj, 2015) (Tang, Yao, Zhang, & Zhang, 2010) (Wang, Ma, & Zhang, 2017) (Xu, et al., 2019) (Zhang, Fu, Jiang, Bao, & Zeng, 2018) (Zhang & Bors, 2019) (Zheng, Li, Zhang, Xie, & Zhong, 2019)**, SVM classifier** (Barbon, Igawa, & Bogaz Zarpelão, 2017) (Chen, Zhang, Chen, Fan, & Gao, 2018)**, SVM with linear kernel** (Hoang & Lim, 2017)**, syntactic and semantic algorithms** (Lee, Hussain, Rivera, & Isroilov, 2018)**, neighborhood based techniques** (Eke, Norman, Shuib, & Nweke, 2019)**, temporal and social probabilistic matrix factorization** (Zarrinkalam, Kahani, & Bagheri, 2019)**, temporal influence correlations (TIC)** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, tensor factorization (TF) models** (Tang, Xu, & Geva, 2014)**, tensor reduction for dimensionality reduction** (Dougnon R. , Fournier-Viger, Lin, & Nkambou, 2016)**, text attention neural network model (TA-NN)** (Li, Yang, Xu, Wang, & Lin, 2019)**, text classification** (Tong, Yao, Wang, & Yang, 2016)**, text feature + HAN** (Zhang, Fu, Jiang, Bao, & Zeng, 2018)**, TF-IDF** (Al Hasan Haldar, Li, Reynolds, Sellis, & Yu, 2019) (Gorrab, Kboubi, Jaffal, Le Grand, & Ghezala, 2017) (Kang, Choi, & Lee, 2019) (Li, Yang, Xu, Wang, & Lin, 2019) (Pla Karidi, Stavrakas, & Vassiliou, 2018) (Yang, Xiao, Tong, Zhang, & Wang, 2015) (Zhang, Fu, Jiang, Bao, & Zeng, 2018)**, TF-IDF features** (Xu, Tadesse, Fei, & Lin, 2019)**, third-parties image classifier (IBM Visual Recognition API)** (Hirt, Kühl, & Satzger, 2019)**, TimeSVD** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, topic model** (Hoang & Lim, 2017) (Kang, Choi, & Lee, 2019) (Ta, Li, Hu, & Feng, 2019) (Wu, Wang, Guo, Zhang, & Chen, 2016)**, TopicMF (matrix factorization)** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, topics model** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, transferable belief model (TBM)** (Pipanmaekaporn & Kamonsantiroj, 2015)**, tree-structured CRF (conditional random field)** (Tang, Yao, Zhang, & Zhang, 2010)**, TrustSVD (latent factor model)** (Zheng, Luo, Sun, Zhang, & Chen, 2018)**, TS-LDA** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, twitterLDA** (Hoang & Lim, 2017)**, two ensembles: SVM with RBF kernel** (Zhuang, Ma, & Yoshikawa, 2017)**, unified discriminative influence model** (Ta, Li, Hu, & Feng, 2019)**, unified discriminative influence probabilistic model** (Huang, Yu, Wang, & Cui, 2015)**, unsupervised method for topic detection** (Nicoletti, Schiaffino, & Godoy, 2013)**, unsupervised multilingual approach** (Bennacer Seghouani, Jipmo, & Quercini, 2019)**, user hierarchical knowledge graphs** (Pla Karidi, Stavrakas, & Vassiliou, 2018)**, user ontology profiling** (Peña, Del Hoyo, Vea-Murguía, González, & Mayo, 2013)**, user similarities by k-NN** (Anand & Mampilli, 2014)**, user-based CF** (Dharia, et al., 2018) (Pla Karidi, Stavrakas, & Vassiliou, 2018) (Wang, Zhong, Yang, & Jing, 2018)**, utility based user profiling mining (UUPM)** (Logesh, Subramaniyaswamy, Vijayakumar, & Li, 2019)**, utility-based systems (UB)** (C C & Mohan, 2019)**, vector space**

model (VSM) (Wang, Ma, & Zhang, 2017)**, weighed-Node2Vec** (Li, et al., 2019)**, Word2Vec** (Kang, Choi, & Lee, 2019) (Tong, Yao, Wang, & Yang, 2016)**, XGBoost** (Chen, Zhang, Chen, Fan, & Gao, 2018) (França, Goya, & Camargo Penteado, 2018)**, XGBoost classifier** (Li, Yang, Xu, Wang, & Lin, 2019)**, XGBoost classifier with linear model** (Xu, Tadesse, Fei, & Lin, 2019)**, XGBoost for content classification** (Tong, Yao, Wang, & Yang, 2016).

Appendix G. Social media user profiling approaches terminology control

affiliation graph model (network structure, probabilistic graphical model), affinity propagation clustering, analysis of variance (ANOVA), applied standardization (clustering), association rule mining, association use mining, asynchronous stochastic gradient algorithm (ASGD), attention mechanism, author topic model, autoencoder-based social recommender system (AESR), autoencoders and perceptrons, averaging models (averaging and stacking models)**, bag of concepts, bag-of-words (BOW)** (bag-of-words (BOW), bag-of-words or topic modeling)**, based on standing ovation model (SOM), Bayesian classification, Bayesian inference, Bayesian networks** (Bayesian network, Bayesian networks (belief networks))**, Bayesian networks and ontologies, Bayesian personalized ranking, Bayesian technique, belief function reasoning, bi-directional gated recurrent unit (biGRU, type of RNN), Bidirectional LSTM (BiLSTM), binary relevance (BR) (user tags prediction), bird flocking, Business Semantics Management, CALGARI and KL-divergence scored, CART tree based model, CBR (case-based reasoning), CENE (network embedding and content), centroid-based classification (text classification), Chinese restaurant process (statistics/probability), classic rank (CLR), classification, classifier chains (CC), clustering, CNN, CNN ResNet-50, CNN to classify images (map to KG), CNN-RNN, co-factorization machines (CoFM), collaborative filtering (activity recommendation engine), Collaborative Filtering (memory-based CF, matrix factorization (SVD, LDA, ALS))** (CF (collaborative filtering), collaborative filtering, Collaborative Filtering (memory-based CF, matrix factorization (SVD, LDA, ALS)), collaborative filtering (model-based and memory-based))**, collaborative filtering rank (CFR), collaborative ontology engineering methods, collective classification, collective naïve bayes, combine semantic context and social network information, combined perceptron with Bayes model, community detection, composite Gaussian Process (GP), compositional recurrent neural network, condensed filter tree (CFT), constrained label propagation, content-based and graph-based features, content-based and preference-based filtering, content-based collaborative filtering** (content-based collaborative filtering, content-based filtering)**, content-based systems, content-based user tag recommendation, context-aware recommender systems (CARS), continuous bag-of-words (CBOW), co-profiling**

algorithm, corr-LDA, Cosine Similarity, Crisp User Profile based Recommendations (CUP), Cross-Media-LDA (CMLDA), DBpedia, DBSCAN, decision tree, decision trees (J48, ADTree, REPTree), deep learning, deep neural network, deep neural networks (bi-GRU layer, hierarchical attention layer, BiRNN, concatenation layer), DeepWalk, demographic systems, dependence distributions, DILIGENT, dimensionality reduction, dimensionality reduction through network embedding paradigm (matrix factorization), discriminative influence model, Dublin Core (Dublin Core, Dublin Core Ontology)**, dynamic user attribute model (DUAM), dynamic user clustering topic model (UCT), dynamic weighted ensemble (DWE), EIUCF, ensemble learning, entropy-based model (EBM), expert systems, explicit semantic analysis, extend deep autoencoder with top-k semantic social information, factor graph model, FCM, Feature Refinement Layer, filtering techniques, fitness buddies recommendation engine, FOAF, formal concept analysis (FCA), FREQ (frequency of tags), frequent pattern mining (FPM) (topic detection), frequent terms (bag of words), friends-based collaborative filtering, fuzzy C-means algorithm for clustering, fuzzy logic (OWA (ordered weighted averaging) operators), gated recurrent neural network (CNN-GRU), Gaussian distribution, Gaussian Mixture Model, Gaussian relational topic model, generalist recommender system kernel (GRSK), generalized matrix factorization, generative influence models, generative relationship influence models, geographic topic models, geographical topic models by utilizing statistical topic models, Gibbs sampling in location-based topic models, GOSPL, GOSPL with D2RQ, gradient boosted decision trees, gradient boosting, graph based approaches (centrality, betweenness), graph embedding, graph embedding algorithms (LINE, PUHE), graph embedding learning, graph partitioning, Graph Theoretic Analysis, graph theory, graph-based (session-based temporal graph), graph-based algorithms** (graph based, graph-based, graph-based algorithms)**, graph-based user tag recommendation, graphical models, GRU structure, HCOME, heterogenous graph embeddings, heuristic approaches (TF, TF-IDF, TI-TextRank, FTF), Hidden Markov Models (HMM)** (Hidden Markov Models (HMM), HMM)**, hierarchical attention network, Hierarchical Attention Transfer Network (HATN), hierarchical Bayesian model, hierarchical clustering, hierarchical convolution neural network (CNN), hierarchical interest graph (from Wikipedia category graph), HMRF (Hidden Markov Random Field), HMRF-KMEANS, hypertext induced topic search (HITS), IBM Watson personality insights, ImageNet/GoogleNet, incremental Bayesian online updates, individual filtering (user preferences), information filtering (IF), injected preferences fusion (IPF), IPE, item-based CF, ITT, k-means (k-means, k-means clustering), k-NN, knowledge graphs, knowledge-based systems (KB), label propagation, labeled-LDA, language models, large-scale information network embedding (LINE)** (large-scale information network embedding (LINE), LINE (large information network embedding))**, latent factor model, Latent Semantic Analysis, Latent Semantic Analysis (LSA) using matrix factorization technique, latent**

semantic hashing, latent SVM (LSVM), LDA, Levenshtein Similarity, linear regression, linear regression (adapted balance winnow algorithm), Linguist Quantifier driven Tag Determination (LQT) , L-LDA, logistic regression, LSTM, LSTM (sentiment classifier), LTPA (local tag propagation), Lucene Clustering, majority vote (majority vote, majority voting)**, map image to knowledge graph entities, markov chains, Markov logic network (MLN), Markov random field, matrix decomposition techniques (specifically non-negative matrix factorization (NMF)), matrix factorization, Maximum Likelihood Estimation, MCMC, memory-based CF, MLP (multi-layer perceptron), mm-LDA, modded SVD (modSVD), model-based CF, model-based recommendation (matrix factorization, modified balanced winnow algorithm (learning a linear classifier), most popular friends (MPF), MRF (Markov Random Field), multi-granularity CNN, multi-modal deep belief network (DBN), multi-modal deep Boltzmann Machines (DBM), multi-model user attribute model (mmUAM), multinominal logistic regression (MLR), Naïve Bayes, naïve Bayes classifier** (naïve Bayes classifier, naïve Bayes for classification)**, naïve Bayes multinominal (NBM), nearest neighbour distribution over ODP (Open Directory Project), neighbourhood-based CF, neighbourhood-based methods** (neighborhood-based methods, similarity-/neighborhood-based, neighborhood based techniques)**, network analysis (graph-based), network embeddings, network representation learning (NRL), neural network model (social convolution attention neural network), neural networks, neural recommendations (neural networks and collaborative filtering), n-grams, N-grams authorship verification (AV), NLP, non-linear (radial basis function) kernel SVM, non-negative matrix factorization (NMF), normalized graphs, OKM, ontologies, ontology based, ontology engineering project, ontology to categorize results, ontology-based recommendations, ontology-based user models, OpenDNS, outliers determined by interquartile ranges (IQR), OWL, Pattern Taxonomy Model (PTM), PCA (clustering), PCFA (principal component factor analysis) (clustering), PDS, PGBN (Poisson Gamma Belief Network, a deep learning topic model), popularity rank (PR), POS (part of speech), probabilistic approaches (Explicit semantic analysis (ESA)), probabilistic framework, probabilistic inference (for user location), probabilistic latent factor models), probabilistic latent semantic analysis (PLSA), probabilistic matrix factorization (PMF)** (probabilistic matrix factorization (PMF), probability matrix factorization (PMF))**, probabilistic model, probabilistic topic model, probability distributions, probability models, QBLDA, QICE, random forest** (random forest, random forest classification)**, Rank based Degree of Feature (RDF), rating-based systems, RDF, recommender systems (social recommendation), relational graph, relational naïve bayes classifier, replicated softmax model, representation learning (feature learning), Restricted Boltzmann Machines (RBMs), RNN, RNN based collaborative filtering, Rocchio, rule-based, rule-based systems, SALSA (stochastic approach for link-structure analysis), semantic methods to recommend friends, semantic relationships, semantic structures,**

semantic technologies for interlinking social websites, semantic trees, semantically enrich user profiles by using association rules, semi-supervised topic model, sentiment analysis (topic models), similarity-based methods (similarity-/neighborhood-based)**, singular value decomposition (SVD)** (singular value decomposition (SVD), SVD)**, SIOC, Smith-Waterman Similarity, social graph, social pertinent walker (SPTW), social tagging system (STS), social-based collaborative filtering (CF), social-based filtering (friends' preferences), socially embedded visual representation learning (SEVIR), Softmax layer, SoRec (social regularization), spectral clustering, spreading activation algorithm (e.g., algorithm over semantic networks), stacked model to do classifier stacking, stacked SVM, stacking and boosting enhanced ensemble, stacking models** (averaging and stacking models, ML models stacking (ensemble), stacking (ensemble))**, statistical analysis, statistical classifier, statistical modelling, stochastic gradient descent classifier, stochastic topic model, supervised topic modelling, SVM, SVM classifier, SVM with linear kernel** (linear kernel SVM, linear SVM, SVM with linear kernel)**, syntactic and semantic algorithms, temporal and social probabilistic matrix factorization, temporal influence correlations (TIC), tensor factorization (TF) models, tensor reduction for dimensionality reduction, text attention neural network model (TA-NN), text classification, text feature + HAN, TF-IDF, TF-IDF features, third-parties image classifier (IBM Visual Recognition API), TimeSVD, topic model, TopicMF (matrix factorization), topics model** (topics model, bag-of-words or topic modeling)**, transferable belief model (TBM), tree-structured CRF (conditional random field), TrustSVD (latent factor model), TS-LDA, twitterLDA, two ensembles: SVM with RBF kernel, unified discriminative influence model, unified discriminative influence probabilistic model, unsupervised method for topic detection, unsupervised multilingual approach, user hierarchical knowledge graphs, user ontology profiling, user similarities by k-NN, user-based CF, utility based user profiling mining (UUPM), utility-based systems (UB), vector space model (VSM), weighed-Node2Vec, Word2Vec, XGBoost** (XGBoost, XGBoost classifier, XGBoost classifier with linear model, XGBoost for content classification)

Appendix H. Social media user profiling approaches

affiliation graph model (network structure, probabilistic graphical model), affinity propagation clustering, analysis of variance (ANOVA), applied standardization (clustering), association rule mining, association use mining, asynchronous stochastic gradient algorithm (ASGD), attention mechanism, author topic model, autoencoder-based social recommender system (AESR), autoencoders and perceptrons, averaging models, bag of concepts, bag-of-words (BOW), based on standing ovation model (SOM), Bayesian classification, Bayesian inference, Bayesian networks, Bayesian networks and ontologies, Bayesian personalized ranking, Bayesian

technique, belief function reasoning, bi-directional gated recurrent unit (biGRU, type of RNN), Bidirectional LSTM (BiLSTM), binary relevance (BR) (user tags prediction), bird flocking, Business Semantics Management, CALGARI and KL-divergence scored, CART tree based model, CBR (case-based reasoning), CENE (network embedding and content), centroid-based classification (text classification), Chinese restaurant process (statistics/probability), classic rank (CLR), classification, classifier chains (CC), clustering, CNN, CNN ResNet-50, CNN to classify images (map to KG), CNN-RNN, co-factorization machines (CoFM), collaborative filtering (activity recommendation engine), Collaborative Filtering (memory-based CF, matrix factorization (SVD, LDA, ALS)), collaborative filtering rank (CFR), collaborative ontology engineering methods, collective classification, collective naïve bayes, combine semantic context and social network information, combined perceptron with Bayes model, community detection, composite Gaussian Process (GP), compositional recurrent neural network, condensed filter tree (CFT), constrained label propagation, content-based and graph-based features, content-based and preference-based filtering, content-based collaborative filtering, content-based systems, content-based user tag recommendation, context-aware recommender systems (CARS), continuous bag-of-words (CBOW), co-profiling algorithm, corr-LDA, Cosine Similarity, Crisp User Profile based Recommendations (CUP), Cross-Media-LDA (CMLDA), DBpedia, DBSCAN, decision tree, decision trees (J48, ADTree, REPTree), deep learning, deep neural network, deep neural networks (bi-GRU layer, hierarchical attention layer, BiRNN, concatenation layer), DeepWalk, demographic systems, dependence distributions, DILIGENT, dimensionality reduction, dimensionality reduction through network embedding paradigm (matrix factorization), discriminative influence model, Dublin Core, dynamic user attribute model (DUAM), dynamic user clustering topic model (UCT), dynamic weighted ensemble (DWE), EIUCF, ensemble learning, entropy-based model (EBM), expert systems, explicit semantic analysis, extend deep autoencoder with top-k semantic social information, factor graph model, FCM, Feature Refinement Layer, filtering techniques, fitness buddies recommendation engine, FOAF, formal concept analysis (FCA), FREQ (frequency of tags), frequent pattern mining (FPM) (topic detection), frequent terms (bag of words), friends-based collaborative filtering, fuzzy C-means algorithm for clustering, fuzzy logic (OWA (ordered weighted averaging) operators), gated recurrent neural network (CNN-GRU), Gaussian distribution, Gaussian Mixture Model, Gaussian relational topic model, generalist recommender system kernel (GRSK), generalized matrix factorization, generative influence models, generative relationship influence models, geographic topic models, geographical topic models by utilizing statistical topic models, Gibbs sampling in location-based topic models, GOSPL, GOSPL with D2RQ, gradient boosted decision trees, gradient boosting, graph based approaches (centrality, betweenness), graph embedding, graph embedding algorithms (LINE, PUHE), graph embedding learning, graph partitioning, Graph Theoretic Analysis, graph theory, graph-based (session-based temporal graph), graph-based algorithms, graph-based user tag recommendation, graphical models, GRU structure,

HCOME, heterogenous graph embeddings, heuristic approaches (TF, TF-IDF, TI-TextRank, FTF), Hidden Markov Models (HMM), hierarchical attention network, Hierarchical Attention Transfer Network (HATN), hierarchical Bayesian model, hierarchical clustering, hierarchical convolution neural network (CNN), hierarchical interest graph (from Wikipedia category graph), HMRF (Hidden Markov Random Field), HMRF-KMEANS, hypertext induced topic search (HITS), IBM Watson personality insights, ImageNet/GoogleNet, incremental Bayesian online updates, individual filtering (user preferences), information filtering (IF), injected preferences fusion (IPF), IPE, item-based CF, ITT, k-means (k-means, k-means clustering), k-NN, knowledge graphs, knowledge-based systems (KB), label propagation, labeled-LDA, language models, large-scale information network embedding (LINE), latent factor model, Latent Semantic Analysis, Latent Semantic Analysis (LSA) using matrix factorization technique, latent semantic hashing, latent SVM (LSVM), LDA, Levenshtein Similarity, linear regression, linear regression (adapted balance winnow algorithm), Linguist Quantifier driven Tag Determination (LQT), L-LDA, logistic regression, LSTM, LSTM (sentiment classifier), LTPA (local tag propagation), Lucene Clustering, majority vote, map image to knowledge graph entities, Markov chains, Markov logic network (MLN), Markov random field, matrix decomposition techniques (specifically non-negative matrix factorization (NMF)), matrix factorization, Maximum Likelihood Estimation, MCMC, memory-based CF, MLP (multi-layer perceptron), mm-LDA, modded SVD (modSVD), model-based CF, model-based recommendation (matrix factorization, modified balanced winnow algorithm (learning a linear classifier), most popular friends (MPF), MRF (Markov Random Field), multi-granularity CNN, multi-modal deep belief network (DBN), multi-modal deep Boltzmann Machines (DBM), multi-model user attribute model (mmUAM), multinominal logistic regression (MLR), Naïve Bayes, naïve Bayes classifier, naïve Bayes multinominal (NBM), nearest neighbor distribution over ODP (Open Directory Project), neighborhood-based CF, neighborhood-based methods, network analysis (graph-based), network embeddings, network representation learning (NRL), neural network model (social convolution attention neural network), neural networks, neural recommendations (neural networks and collaborative filtering), n-grams, N-grams authorship verification (AV), NLP, non-linear (radial basis function) kernel SVM, non-negative matrix factorization (NMF), normalized graphs, OKM, ontologies, ontology based, ontology engineering project, ontology to categorize results, ontology-based recommendations, ontology-based user models, OpenDNS, outliers determined by interquartile ranges (IQR), OWL, Pattern Taxonomy Model (PTM), PCA (clustering), PCFA (principal component factor analysis) (clustering), PDS, PGBN (Poisson Gamma Belief Network, a deep learning topic model), popularity rank (PR), POS (part of speech), probabilistic approaches (Explicit semantic analysis (ESA)), probabilistic framework, probabilistic inference (for user location), probabilistic latent factor models), probabilistic latent semantic analysis (PLSA), probabilistic matrix factorization (PMF), probabilistic model, probabilistic topic model, probability distributions, probability models, QBLDA, QICE, random

forest, Rank based Degree of Feature (RDF), rating-based systems, RDF, recommender systems (social recommendation), relational graph, relational naïve Bayes classifier, replicated softmax model, representation learning (feature learning), Restricted Boltzmann Machines (RBMs), RNN, RNN based collaborative filtering, Rocchio, rule-based, rule-based systems, SALSA (stochastic approach for link-structure analysis), semantic methods to recommend friends, semantic relationships, semantic structures, semantic technologies for interlinking social websites, semantic trees, semantically enrich user profiles by using association rules, semi-supervised topic model, sentiment analysis (topic models), similarity-based methods, singular value decomposition (SVD), SIOC, Smith-Waterman Similarity, social graph, social pertinent walker (SPTW), social tagging system (STS), social-based collaborative filtering (CF), social-based filtering (friends' preferences), socially embedded visual representation learning (SEVIR), Softmax layer, SoRec (social regularization), spectral clustering, spreading activation algorithm (e.g., algorithm over semantic networks), stacked model to do classifier stacking, stacked SVM, stacking and boosting enhanced ensemble, stacking models, statistical analysis, statistical classifier, statistical modeling, stochastic gradient descent classifier, stochastic topic model, supervised topic modeling, SVM, SVM classifier, SVM with linear kernel, syntactic and semantic algorithms, temporal and social probabilistic matrix factorization, temporal influence correlations (TIC), tensor factorization (TF) models, tensor reduction for dimensionality reduction, text attention neural network model (TA-NN), text classification, text feature + HAN, TF-IDF, TF-IDF features, third-parties image classifier (IBM Visual Recognition API), TimeSVD, topic model, TopicMF (matrix factorization), topics model, transferable belief model (TBM), tree-structured CRF (conditional random field), TrustSVD (latent factor model), TS-LDA, twitterLDA, two ensembles: SVM with RBF kernel, unified discriminative influence model, unified discriminative influence probabilistic model, unsupervised method for topic detection, unsupervised multilingual approach, user hierarchical knowledge graphs, user ontology profiling, user similarities by k-NN, user-based CF, utility based user profiling mining (UUPM), utility-based systems (UB), vector space model (VSM), weighed-Node2Vec, Word2Vec, XGBoost.

Appendix I. Explainability techniques SLR search results

Title	Author(s)	Year
Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges	Preece, Alun	2018
Explainability in human–agent systems	Rosenfeld, Avi; Richardson, Ariella	2019

Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending	Ariza, Miller; Arroyo, Javier; Caparrini, Antonio; Segovia, Maria-Jesus	2020
Mining Semantic Knowledge Graphs to Add Explainability to Black Box Recommender Systems	Alshammari, Mohammed; Nasraoui, Olfa; Sanders, Scott	2019
Improving Explainability of Recommendation System by Multi-sided Tensor Factorization	Hong, Minsung; Akerkar, Rajendra; Jung, Jason J	2019
Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models	Bikmukhametov, Timur; Jäschke, Johannes	2020
Definitions, methods, and applications in interpretable machine learning	Murdoch, W James; Singh, Chandan; Kumbier, Karl; Abbasi-Asl, Reza; Yu, Bin	2019
Machine Learning Interpretability: A Survey on Methods and Metrics	Diogo V. Carvalho; Eduardo M. Pereira; Jaime S. Cardoso	2019
Explainable Deep Learning Models in Medical Image Analysis	Amitojdeep Singh; Sourya Sengupta; Vasudevan Lakshminarayanan	2020
Understanding the decisions of CNNs: An in-model approach	Rio-Torto, Isabel; Fernandes, Kelwin; Teixeira, Luís F	2020
Relational social recommendation: Application to the academic domain	Amal, Saeed; Tsai, Chun-Hua; Brusilovsky, Peter; Kuflik, Tsvi; Minkov, Einat	2019
Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification	Böhle, Moritz; Eitel, Fabian; Weygandt, Martin; Ritter, Kerstin	2019
Explaining Support Vector Machines: A Color Based Nomogram	Vanya Van Belle; Ben Van Calster; Sabine Van Huffel; Johan A K Suykens; Paulo Lisboa	2016

Band Selection via Explanations From Convolutional Neural Networks	Karlsson, Isak; Rebane, Jonathan; Papapetrou, Panagiotis; Gionis, Aristides	2019
"What is relevant in a text document?": An interpretable machine learning approach	Arras, Leila; Horn, Franziska; Montavon, Grégoire; Müller, Klaus-Robert; Samek, Wojciech	2017
Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome	Esra Zihni; Vince Istvan Madai; Michelle Livne; Ivana Galinovic; Ahmed A Khalil; Jochen B Fiebach; Dietmar Frey	2020
Explainable Machine Learning for Scientific Insights and Discoveries	Roscher, Ribana; Bohn, Bastian; Duarte, Marco F; Garcke, Jochen	2020
explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning	Spinner, Thilo; Schlegel, Udo; Schafer, Hanna; El-Assady, Mennatallah	2020
Explanations for Temporal Recommendations	Bharadhwaj, Homanga; Joshi, Shruti	2018
Explainable Machine Learning Framework for Image Classification Problems: Case Study on Glioma Cancer Prediction	Emmanuel Pintelas; Meletis Liaskos; Ioannis E. Livieris; Sotiris Kotsiantis; Panagiotis Pintelas	2020
Reliable and explainable machine-learning methods for accelerated material discovery	Bhavya Kailkhura; Brian Gallagher; Sookyung Kim; Anna Hiszpanski; T. Yong-Jin Han	2019
A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability	Emmanuel Pintelas; Ioannis E. Livieris; Panagiotis Pintelas	2020
Towards a demystification of the Black Box – explainable AI and legal ramifications	Käde, Lisa; Von Maltzan, Stephanie	2019
DARPA's Explainable Artificial Intelligence Program	Gunning, David; Aha, David	2019
Defining Explainable AI for Requirements Analysis	Sheh, Raymond; Monteath, Isaac	2018
Explainable AI: from black box to glass box	Rai Arun	2020

Explainable recommendation with fusion of aspect information	Hou, Yunfeng; Yang, Ning; Wu, Yi; Yu, Philip	2019
User Evaluations on Sentiment-based Recommendation Explanations	Chen, Li; Yan, Dongning; Wang, Feng	2019
Concept attribution: Explaining CNN decisions to physicians	M, Graziani; V, Andrearczyk; S, Marchand-Maillet; H, Müller	2020
Interpreting Recurrent Neural Networks Behaviour via Excitable Network Attractors	Ceni, A; Ashwin, P; Livi, L	2020
Explaining clusterings of process instances	De Koninck, Pieter; De Weerdt, Jochen; vanden Broucke, Seppe	2017
A Survey of Methods for Explaining Black Box Models	Guidotti, Riccardo; Monreale, Anna; Ruggieri, Salvatore; Turini, Franco; Giannotti, Fosca; Pedreschi, Dino	2019
MMALFM: Explainable Recommendation by Leveraging Reviews and Images	Cheng, Zhiyong; Chang, Xiaojun; Zhu, Lei; Kanjirathinkal, Rose; Kankanhalli, Mohan	2019
Techniques for interpretable machine learning	Du, Mengnan; Liu, Ninghao; Hu, Xia	2019

Table 16. Explainability of social media user profiling approaches SLR search results

Appendix J. Explainability techniques extracted terms

1Rule (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, 3-Level Explanation** (Gunning & Aha, 2019)**, Acceptance Testing** (Gunning & Aha, 2019)**, Accumulated Local Effects Plot** (Carvalho, Pereira, & Cardoso, 2019)**, activation maps** (Singh, Sengupta, & Lakshminarayanan, 2020)**, agnostic explanators** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, AIF360** (Käde & Von Maltzan, 2019)**, algorithmic transparency** (Carvalho, Pereira, & Cardoso, 2019) (Preece, 2018) (Roscher, Bohn, Duarte, & Garcke, 2020)**, analysis of layers of a 3D-CNN using Gaussian mixture model (GMM) and binary encoding of training and test images based on their GMM components for returning similar 3D images** (Singh, Sengupta, & Lakshminarayanan, 2020)**, anchors** (Carvalho, Pereira, & Cardoso, 2019) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**,**

approximate an interpretable model for the black-box model (Arun, 2020)**, Argumentation and Pedagogy** (Gunning & Aha, 2019)**, argumentation theory based** (Gunning & Aha, 2019)**, Aspect-based Matrix Factorization model (AMF)** (Hou, Yang, Wu, & Yu, 2019)**, attention based model** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, attention heatmaps** (Roscher, Bohn, Duarte, & Garcke, 2020)**, attention maps** (Singh, Sengupta, & Lakshminarayanan, 2020)**, attention mask weights** (Roscher, Bohn, Duarte, & Garcke, 2020)**, attention mechanism** (Du, Liu, & Hu, 2019) (Arun, 2020)**, attention modules** (Roscher, Bohn, Duarte, & Garcke, 2020)**, AttentiveChrome NN** (Roscher, Bohn, Duarte, & Garcke, 2020)**, attribute explanations** (Sheh & Monteath, 2018)**, attribute identity explanations** (Sheh & Monteath, 2018)**, attribute use explanations** (Sheh & Monteath, 2018)**, attribution maps** (Singh, Sengupta, & Lakshminarayanan, 2020)**, attributions** (Singh, Sengupta, & Lakshminarayanan, 2020)**, auditing** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, auxiliary criteria** (Preece, 2018)**, auxiliary data** (Bharadhwaj & Joshi, 2018)**, auxiliary network** (Roscher, Bohn, Duarte, & Garcke, 2020)**, backpropagation of the gradients** (M, V, S, & H, 2020)**, back-propagation-based methods** (Du, Liu, & Hu, 2019)**, backward propagation methods** (Zihni, et al., 2020)**, Bayesian Case Model (BCM)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Rosenfeld & Richardson, 2019)**, Bayesian Rule Lists (BRL)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Bayesian Teaching** (Gunning & Aha, 2019)**, BreakDown** (Carvalho, Pereira, & Cardoso, 2019)**, CAM** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, capsule network** (Du, Liu, & Hu, 2019)**, case-based explanations** (Chen, Yan, & Wang, 2019)**, case-based reasoning (case-based explanations by example)** (Preece, 2018)**, causal explanations** (Preece, 2018)**, causal models to explain learning (CAMEL) approach** (Gunning & Aha, 2019)**, CAV.** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, CENTAUR** (Preece, 2018)**, CFS** (Rosenfeld & Richardson, 2019)**, class activation mapping** (Rosenfeld & Richardson, 2019)**, class maps** (Preece, 2018)**, CLEAR (CLass-Enhanced Attentive Response)** (Preece, 2018)**, clustering methods** (Carvalho, Pereira, & Cardoso, 2019)**, coalition game theory to evaluate the effect of combinations of features** (Rosenfeld & Richardson, 2019)**, coalitional game theory based** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, co-clustering approach to gain explainability in a user-item bipartite network** (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019)**, collaborative filtering method using a tensor modeled by considering the 5Ws with explanations based on template** (Hong, Akerkar, & Jung, 2019)**, collaborative-based explanations** (Chen, Yan, & Wang, 2019)**, color based nomogram** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, combination of genetic algorithms with decision trees or rules** (Rosenfeld & Richardson, 2019)**, combination with other machine learning methods** (Singh, Sengupta, & Lakshminarayanan, 2020)**, combine model compression with dimension reduction** (Carvalho, Pereira, & Cardoso, 2019)**, combine physics-based models with machine**

learning (Bikmukhametov & Jäschke, 2020)**, Combined Multiple Model (CCM)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, community tags to explain recommendations** (Alshammari, Nasraoui, & Sanders, 2019)**, complemental examples** (Carvalho, Pereira, & Cardoso, 2019)**, complementary examples** (Rio-Torto, Fernandes, & Teixeira, 2020)**, Compound Critiques** (Chen, Yan, & Wang, 2019)**, Concept Activation Vectors (TCAV)** (Carvalho, Pereira, & Cardoso, 2019)**, concept vectors** (Singh, Sengupta, & Lakshminarayanan, 2020)**, concept-based explanations** (M, V, S, & H, 2020)**, conceptual clustering** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, Confident Decision Tree (CDT)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Conj Rules** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, constraint programming for converting linear SVM (and other hyperplane-based linear classifiers) into a set of non overlapping and interpretable rules** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, content-based explanations** (Chen, Yan, & Wang, 2019) (Hong, Akerkar, & Jung, 2019)**, contextual decomposition explanation penalization** (Roscher, Bohn, Duarte, & Garcke, 2020)**, counterfactual explanations** (Carvalho, Pereira, & Cardoso, 2019) (Käde & Von Maltzan, 2019) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Pintelas, Livieris, & Pintelas, 2020)**, CPAR (Classification based on Predictive Association Rules)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, cross-cluster model-log alignment for identifying differences between clusters** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, CTR** (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)**, data points** (Carvalho, Pereira, & Cardoso, 2019) (Pintelas, Livieris, & Pintelas, 2020)**, data visualization methods** (Carvalho, Pereira, & Cardoso, 2019)**, data-dependent** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, dataset-level** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, Dead Weight** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, Decision Diagrams** (Gunning & Aha, 2019)**, decision rules** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Rosenfeld & Richardson, 2019) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, decision trees** (Carvalho, Pereira, & Cardoso, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020) (Roscher, Bohn, Duarte, & Garcke, 2020) (Rosenfeld & Richardson, 2019)**, decomposability** (Carvalho, Pereira, & Cardoso, 2019)**, decompositional rule extraction** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, DeConvNet** (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, deconvolution** (Rio-Torto, Fernandes, & Teixeira, 2020) (Zihni, et al., 2020) (Böhle, Eitel, Weygandt, & Ritter, 2019)**, deconvolutional networks** (Gunning & Aha, 2019)**, DecText** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Deep Attention-Based Representations for Explanation/Explainable Generative Adversarial Networks (DARE/X-GANS)** (Gunning & Aha, 2019)**, Deep Learning Important FeaTures (DeepLIFT)** (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**,**

deep explanation (Gunning & Aha, 2019)**, Deep Hierarchical Generative Models** (Singh, Sengupta, & Lakshminarayanan, 2020)**, Deep SHapley Additive exPlanations** (Singh, Sengupta, & Lakshminarayanan, 2020)**, deep Taylor decomposition** (Arras, Horn, Montavon, Müller, & Samek, 2017) (Zihni, et al., 2020)**, deep tensor networks** (Roscher, Bohn, Duarte, & Garcke, 2020)**, DeepDreams** (Singh, Sengupta, & Lakshminarayanan, 2020)**, DeepExplain** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, DeepTaylor** (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, DeepTune** (Roscher, Bohn, Duarte, & Garcke, 2020)**, descriptive statistics** (Carvalho, Pereira, & Cardoso, 2019)**, design transparency** (Roscher, Bohn, Duarte, & Garcke, 2020)**, diagnostic sentence** (Singh, Sengupta, & Lakshminarayanan, 2020)**, domain constraints** (Singh, Sengupta, & Lakshminarayanan, 2020)**, domain knowledge supported interpretation** (Zihni, et al., 2020)**, domain-based feature engineering** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, domain-dependent** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, dual neural network system** (Preece, 2018)**, dual system approach** (Preece, 2018)**, EG** (Singh, Sengupta, & Lakshminarayanan, 2020)**, e-LRP** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, enforce sparsity terms** (Du, Liu, & Hu, 2019)**, ERBM** (Bharadhwaj & Joshi, 2018)**, example-based explanation** (Gunning & Aha, 2019) (Käde & Von Maltzan, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020)**, examples** (Preece, 2018)**, excitable network attractors (ENAs)** (Ceni, Ashwin, & Livi, 2020)**, expert knowledge** (Singh, Sengupta, & Lakshminarayanan, 2020)**, expert-determined features relevance** (Preece, 2018)**, explain the recommendation process** (Chen, Yan, & Wang, 2019)**, Explainable Expert Systems (EES) project** (Preece, 2018)**, Explainable Matrix Factorization (EMF)** (Alshammari, Nasraoui, & Sanders, 2019) (Bharadhwaj & Joshi, 2018)**, explainable question answering system (EQUAS)** (Gunning & Aha, 2019)**, explainable recommendation** (Hou, Yang, Wu, & Yu, 2019)**, explainer** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, explanation in terms of input variables** (Arras, Horn, Montavon, Müller, & Samek, 2017)**, explanation maps** (Roscher, Bohn, Duarte, & Garcke, 2020)**, explanation-based learning** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, explanations for a hybrid recommender system** (Chen, Yan, & Wang, 2019)**, explanations in time-series recommendation** (Bharadhwaj & Joshi, 2018)**, explicit factor model** (Hong, Akerkar, & Jung, 2019)**, explicitly capturing and displaying the interactions learned by a neural network** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, Factorized Latent Aspect ModEl (FLAME) combining collaborative filtering and opinion mining** (Hong, Akerkar, & Jung, 2019)**, feature analysis** (Rosenfeld & Richardson, 2019)**, feature attribution** (Carvalho, Pereira, & Cardoso, 2019)**, feature engineering** (Bikmukhametov & Jäschke, 2020)**, feature extraction and explanation extraction framework** (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020)**, feature importance** (Arun, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Du, Liu, & Hu, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Roscher, Bohn, Duarte, & Garcke, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner,

Schlegel, Schafer, & El-Assady, 2020) (Zihni, et al., 2020)**, feature importance analysis** (Bikmukhametov & Jäschke, 2020)**, feature importance and ranking** (Zihni, et al., 2020)**, feature importance scores** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, feature influence methods** (Käde & Von Maltzan, 2019)**, Feature Interaction** (Carvalho, Pereira, & Cardoso, 2019)**, feature sentiments** (Chen, Yan, & Wang, 2019)**, feature summary** (Carvalho, Pereira, & Cardoso, 2019)**, feature summary statistic** (Pintelas, Livieris, & Pintelas, 2020)**, feature summary visualization** (Pintelas, Livieris, & Pintelas, 2020)**, filters** (Rosenfeld & Richardson, 2019)**, FINE (feature importance in nonlinear embeddings)** (Roscher, Bohn, Duarte, & Garcke, 2020)**, Forest Floor** (Käde & Von Maltzan, 2019)**, four-order tensor to model users** (Hong, Akerkar, & Jung, 2019)**, FRL** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Gaussian Process Classification (GDP)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Generalized Additive Models (GAMs)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, generative modeling** (Singh, Sengupta, & Lakshminarayanan, 2020)**, genetic programming** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, global** (Singh, Sengupta, & Lakshminarayanan, 2020)**, global explanation methods** (De Koninck, De Weerdt, & vanden Broucke, 2017) (Du, Liu, & Hu, 2019)**, global in scope** (Arun, 2020)**, global interpretability** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, global level of explainability** (Käde & Von Maltzan, 2019)**, global model interpretability** (Carvalho, Pereira, & Cardoso, 2019)**, GMM and atlas** (Singh, Sengupta, & Lakshminarayanan, 2020)**, GNNExplainer** (Singh, Sengupta, & Lakshminarayanan, 2020)**, GoldenEye** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, GPDT** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, grad*input** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, Grad-CAM** (Carvalho, Pereira, & Cardoso, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, GradHM+AS** (Karlsson, Rebane, Papapetrou, & Gionis, 2019)**, GradHM+TS** (Karlsson, Rebane, Papapetrou, & Gionis, 2019)**, Gradient** (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, gradient methods** (Böhle, Eitel, Weygandt, & Ritter, 2019)**, Gradient weighted class activation mapping (GradCAM)** (Böhle, Eitel, Weygandt, & Ritter, 2019) (Karlsson, Rebane, Papapetrou, & Gionis, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020)**, Gradient x input** (Singh, Sengupta, & Lakshminarayanan, 2020)**, Gradient*I/P** (Singh, Sengupta, & Lakshminarayanan, 2020)**, gradient-based approaches** (M, V, S, & H, 2020)**, gradient-based methods** (Zihni, et al., 2020)**, gradient-weighted heatmap (GradHM)** (Karlsson, Rebane, Papapetrou, & Gionis, 2019)**, graph-based recommendation approach** (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019)**, graphing the functional relationship between the predicted response and the feature for individual observations** (Rosenfeld & Richardson, 2019)**, G-REX** (Guidotti, et al., A Survey of Methods for Explaining Black Box

Models, 2019)**, GroupINN** (Roscher, Bohn, Duarte, & Garcke, 2020)**, GSInquire** (Singh, Sengupta, & Lakshminarayanan, 2020)**, guided backpropagation** (Böhle, Eitel, Weygandt, & Ritter, 2019) (Carvalho, Pereira, & Cardoso, 2019) (Karlsson, Rebane, Papapetrou, & Gionis, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020) (Zihni, et al., 2020)**, Guided Grad-CAM** (Rio-Torto, Fernandes, & Teixeira, 2020)**, Guided-GradHM** (Karlsson, Rebane, Papapetrou, & Gionis, 2019)**, Guided-GradHM+AS** (Karlsson, Rebane, Papapetrou, & Gionis, 2019)**, Guided-GradHM+TS** (Karlsson, Rebane, Papapetrou, & Gionis, 2019)**, heat maps** (Preece, 2018)**, heatmaps** (Roscher, Bohn, Duarte, & Garcke, 2020)**, heatmaps of salient regions** (M, V, S, & H, 2020)**, HFT** (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)**, HIN technique** (Alshammari, Nasraoui, & Sanders, 2019)**, histogram** (Chen, Yan, & Wang, 2019)**, HistoTrend** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, hybrid approach using collaborative and content-based filtering techniques** (Alshammari, Nasraoui, & Sanders, 2019)**, identification of prototypes** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, identify prototypes** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, image reconstruction** (Käde & Von Maltzan, 2019)**, individual conditional expectation** (Arun, 2020) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Carvalho, Pereira, & Cardoso, 2019)**, inductive logic programming** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, Influence Functions** (Carvalho, Pereira, & Cardoso, 2019) (Preece, 2018)**, influential data points** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, Info Flow** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, Information Plane** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Information Plane visualization** (Rosenfeld & Richardson, 2019)**, in-model** (Carvalho, Pereira, & Cardoso, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020)**, in-model joint architecture from explainer and classifier** (Rio-Torto, Fernandes, & Teixeira, 2020)**, iNNvestigate** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, Input × Gradient** (Rio-Torto, Fernandes, & Teixeira, 2020)**, input variables influence** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, instance-level explanations** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, integrate explanations into Matrix Factorization** (Bharadhwaj & Joshi, 2018)**, Integrated Grad** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, integrated gradients** (Carvalho, Pereira, & Cardoso, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020)**, integration of feature interaction and tree interpretation functionalities into Random Forest program code** (Käde & Von Maltzan, 2019)**, interaction and feature importances** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, Interactive Training** (Gunning & Aha, 2019)**, Interactive Visualization** (Gunning & Aha, 2019)**, interpretability constraints** (Du, Liu, & Hu, 2019)**, interpretability constraints into the structure of the model** (Arun, 2020)**, interpretable convolutional neural networks** (Du, Liu, & Hu, 2019)**, Interpretable Decision Sets (IDS)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, interpretable**

mimic learning (Pintelas, Livieris, & Pintelas, 2020)**, interpretable model extraction** (Du, Liu, & Hu, 2019)**, interpretable models** (Gunning & Aha, 2019)**, InterpretML-Framework** (Käde & Von Maltzan, 2019)**, inTrees** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, intrinsic** (Carvalho, Pereira, & Cardoso, 2019)**, intrinsic explainability** (Singh, Sengupta, & Lakshminarayanan, 2020)**, intrinsic explanation** (Du, Liu, & Hu, 2019)**, intrinsic interpretability** (Pintelas, Livieris, & Pintelas, 2020)**, intrinsic interpretable Grey-Box ensemble model** (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020)**, intrinsic methods** (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020)**, Introspective explanations** (Sheh & Monteath, 2018)**, investigation of deep representations** (Du, Liu, & Hu, 2019)**, items** (Hong, Akerkar, & Jung, 2019)**, ITLFM** (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)**, justification explanations** (Sheh & Monteath, 2018)**, justify why the recommendation might be good for a user** (Chen, Yan, & Wang, 2019)**, keywords and neighbours and ratings** (Chen, Yan, & Wang, 2019)**, keywords or user-tags based explanations** (Chen, Yan, & Wang, 2019)**, k-means** (Carvalho, Pereira, & Cardoso, 2019)**, kNN** (Singh, Sengupta, & Lakshminarayanan, 2020)**, knowledge distillation** (Carvalho, Pereira, & Cardoso, 2019)**, knowledge extraction** (Käde & Von Maltzan, 2019)**, knowledge-based explanations** (Chen, Yan, & Wang, 2019)**, Layer wise relevance propagation (LRP)** (Arras, Horn, Montavon, Müller, & Samek, 2017) (Böhle, Eitel, Weygandt, & Ritter, 2019) (Käde & Von Maltzan, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020) (Roscher, Bohn, Duarte, & Garcke, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Zihni, et al., 2020)**, leverage topic models to discover explainable latent factors in matrix factorization** (Hou, Yang, Wu, & Yu, 2019)**, linear dimensionality reduction** (Roscher, Bohn, Duarte, & Garcke, 2020)**, linear models** (Roscher, Bohn, Duarte, & Garcke, 2020)**, linked data** (Alshammari, Nasraoui, & Sanders, 2019)**, local** (Singh, Sengupta, & Lakshminarayanan, 2020)**, local approximation-based explanation** (Du, Liu, & Hu, 2019)**, local explanation** (Du, Liu, & Hu, 2019)**, local in scope** (Arun, 2020)**, local interpretability** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Local Interpretable Model-Agnostic Explanation (LIME)** (Ariza, Arroyo, Caparrini, & Segovia, 2020) (Arun, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Käde & Von Maltzan, 2019) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Pintelas, Livieris, & Pintelas, 2020) (Preece, 2018) (Rio-Torto, Fernandes, & Teixeira, 2020) (Roscher, Bohn, Duarte, & Garcke, 2020) (Rosenfeld & Richardson, 2019) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, local interpretations** (Rosenfeld & Richardson, 2019)**, local level of explainability** (Käde & Von Maltzan, 2019)**, local model interpretability** (Carvalho, Pereira, & Cardoso, 2019)**, Local Surrogate Model** (Carvalho, Pereira, & Cardoso, 2019)**, local-level** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, LSTM** (Singh, Sengupta, & Lakshminarayanan, 2020)**, Mapping between image to reports** (Singh, Sengupta, & Lakshminarayanan, 2020)**, mask perturbation** (Du, Liu, & Hu, 2019)**, maximum mean**

discrepancy (Rosenfeld & Richardson, 2019)**, MDNet** (Singh, Sengupta, & Lakshminarayanan, 2020)**, mimic learning** (Du, Liu, & Hu, 2019)**, MinMax** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, MMALFM** (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)**, MMD-critic (Maximum Mean Discrepancy)** (Carvalho, Pereira, & Cardoso, 2019)**, model agnostic explanations** (Sheh & Monteath, 2018)**, model coefficients for logistic regression** (Zihni, et al., 2020)**, model compression** (Carvalho, Pereira, & Cardoso, 2019)**, Model Explanation System (MES)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, model explanations** (Sheh & Monteath, 2018)**, model induction** (Gunning & Aha, 2019)**, model internals** (Carvalho, Pereira, & Cardoso, 2019) (Pintelas, Livieris, & Pintelas, 2020)**, model modularity** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, model simulatability** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, model sparsity** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, model tool** (Rosenfeld & Richardson, 2019)**, model-agnostic** (Arun, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Käde & Von Maltzan, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Roscher, Bohn, Duarte, & Garcke, 2020)**, model-based** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, model-based feature engineering** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, model-specific** (Arun, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, monotonic constraints** (Singh, Sengupta, & Lakshminarayanan, 2020)**, MoviExplain** (Alshammari, Nasraoui, & Sanders, 2019)**, multivariate filters** (Rosenfeld & Richardson, 2019)**, MYCIN** (Preece, 2018) (Rosenfeld & Richardson, 2019)**, Narrative Generation** (Gunning & Aha, 2019)**, natural language caption generation** (Preece, 2018)**, nearest neighbors** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, Neighbor ratings** (Chen, Yan, & Wang, 2019)**, neighborhood of an instance** (Arun, 2020)**, neighborhood technique based on cosine similarity** (Alshammari, Nasraoui, & Sanders, 2019)**, neighbourhood based Collaborative Filtering (CF)** (Bharadhwaj & Joshi, 2018)**, neighbourhood style explanation** (Bharadhwaj & Joshi, 2018)**, NEOMYCIN** (Preece, 2018) (Rosenfeld & Richardson, 2019)**, network propagation technique based on deconvolutions to reconstruct input image patterns that are linked to a particular feature map activation or prediction** (Arras, Horn, Montavon, Müller, & Samek, 2017)**, neural activation visualization** (Rosenfeld & Richardson, 2019)**, neural interaction detection** (Roscher, Bohn, Duarte, & Garcke, 2020)**, Neural Interpretation Diagram (NID)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, neural rating and tips generation** (Hong, Akerkar, & Jung, 2019)**, Neurons Activation (NA)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, NL explanations** (Preece, 2018)**, NL generation** (Gunning & Aha, 2019)**, NL justifications** (Gunning & Aha, 2019)**, Node-Link Vis** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, nonlinear dimensionality reduction** (Roscher, Bohn, Duarte, & Garcke, 2020)**, oblique tree sparse additive models (OT-SpAMs)** (Guidotti, et al., A Survey of Methods

for Explaining Black Box Models, 2019)**, occlusion** (Böhle, Eitel, Weygandt, & Ritter, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, One-variable-at-a-Time approach** (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020)**, optimal selection of teaching examples** (Gunning & Aha, 2019)**, Orthogonal Projection of Input Attributes (OPIA)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, outcome tool** (Rosenfeld & Richardson, 2019)**, PALM** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, partial dependence bars** (Rosenfeld & Richardson, 2019)**, Partial Dependence Plot** (Arun, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Partial Dependency Plots (PDP)** (Rosenfeld & Richardson, 2019)**, Pattern Attribution** (Singh, Sengupta, & Lakshminarayanan, 2020)**, PatternNet** (Singh, Sengupta, & Lakshminarayanan, 2020)**, pedagogical rule extraction** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, permutation feature importance** (Du, Liu, & Hu, 2019) (Pintelas, Livieris, & Pintelas, 2020)**, perturbation** (Singh, Sengupta, & Lakshminarayanan, 2020)**, perturbation-based explanation** (Du, Liu, & Hu, 2019)**, post hoc application of supervised learning with support vector machines** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, post hoc interpretability** (Roscher, Bohn, Duarte, & Garcke, 2020)**, post-hoc** (Carvalho, Pereira, & Cardoso, 2019) (M, V, S, & H, 2020) (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, post-hoc analysis** (Rosenfeld & Richardson, 2019)**, post-hoc explanation** (Du, Liu, & Hu, 2019)**, post-hoc interpretability** (Arun, 2020) (Pintelas, Livieris, & Pintelas, 2020)**, post-hoc mechanism to generate explanations in recommender systems** (Alshammari, Nasraoui, & Sanders, 2019)**, post-hoc methods** (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020)**, Post-Hoc Rationalisation** (Sheh & Monteath, 2018)**, post-model** (Carvalho, Pereira, & Cardoso, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020)**, prediction-level** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, Preference-based Organization (Pref-ORG)** (Chen, Yan, & Wang, 2019)**, pre-model** (Carvalho, Pereira, & Cardoso, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020)**, Principal Component Analysis** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016) (Carvalho, Pereira, & Cardoso, 2019) (Roscher, Bohn, Duarte, & Garcke, 2020) (Rosenfeld & Richardson, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020)**, probabilistic generative model** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Prospector** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, prototype** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, prototype analysis** (Rosenfeld & Richardson, 2019)**, prototype and criticism generation** (Rio-Torto, Fernandes, & Teixeira, 2020)**, prototype generation by greedy approach** (Rosenfeld & Richardson, 2019)**, prototype generation by LP relaxation with randomized rounding** (Rosenfeld & Richardson, 2019)**, Prototype Selection (PS)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Roscher, Bohn, Duarte, & Garcke, 2020)**, Prototypes and Criticisms** (Carvalho, Pereira, & Cardoso, 2019)**, proxy models** (Roscher, Bohn, Duarte, &

Garcke, 2020)**, Quantitative Input Influence (QII)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Query evidence that explains DNN decisions** (Gunning & Aha, 2019)**, random perturbations** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, random sampling based** (Arras, Horn, Montavon, Müller, & Samek, 2017)**, Rational explanations** (Gunning & Aha, 2019)**, rationales as part of the learning process** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, RBLT** (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)**, reducing complex NN** (Roscher, Bohn, Duarte, & Garcke, 2020)**, Reflexive explanations** (Gunning & Aha, 2019)**, REFNE** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, relational connecting paths** (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019)**, relevance of input pixels** (M, V, S, & H, 2020)**, relevance scores** (Roscher, Bohn, Duarte, & Garcke, 2020)**, RETAIN (REverse Time AttentIoN)** (Roscher, Bohn, Duarte, & Garcke, 2020)**, reverse engineering** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Roscher, Bohn, Duarte, & Garcke, 2020)**, reviews with ratings to enhance the explainability of matrix factorization** (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)**, RippleNet** (Alshammari, Nasraoui, & Sanders, 2019)**, RMR** (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)**, root causes of misclassifications** (Preece, 2018)**, root causes of process model differences** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, Rule Based Explanator** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, rule extraction** (De Koninck, De Weerdt, & vanden Broucke, 2017) (Roscher, Bohn, Duarte, & Garcke, 2020)**, Rule Set** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, rule-based methods** (Singh, Sengupta, & Lakshminarayanan, 2020)**, rule-based methods for recommender systems** (Bharadhwaj & Joshi, 2018)**, rule-based segmentation** (Singh, Sengupta, & Lakshminarayanan, 2020)**, rule-based segmentation followed by a perturbation analysis** (Singh, Sengupta, & Lakshminarayanan, 2020)**, RxREN** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, saliency** (Rio-Torto, Fernandes, & Teixeira, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Zihni, et al., 2020)**, saliency heatmaps** (Arras, Horn, Montavon, Müller, & Samek, 2017)**, saliency maps** (Böhle, Eitel, Weygandt, & Ritter, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Rosenfeld & Richardson, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, Saliency Masks** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Roscher, Bohn, Duarte, & Garcke, 2020) (Rosenfeld & Richardson, 2019)**, salient (highest weighted or most predictive) text features or fragments** (Preece, 2018)**, salient examples** (Preece, 2018)**, salient part of the images** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, salient sentences from text documents using loss gradient magnitudes** (Arras, Horn, Montavon, Müller, & Samek, 2017)**, salient structures within images related to a specific class by computing the corresponding prediction score derivative with respect to the input image** (Arras, Horn, Montavon, Müller,

& Samek, 2017)**, Saturated Weight** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, SAUNet** (Singh, Sengupta, & Lakshminarayanan, 2020)**, search for explanations of clusters of process instances (SECPI)** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, second deep network that generates explanations** (Gunning & Aha, 2019)**, self-organizing maps** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, self-training Grey-Box model** (Pintelas, Livieris, & Pintelas, 2020)**, semantic distance (LDSD) algorithm and DBpedia based recommendation system** (Alshammari, Nasraoui, & Sanders, 2019)**, semantic meaningfulness constraints** (Arun, 2020)**, semantic monotonicity constraints** (Du, Liu, & Hu, 2019)**, semantic property values to explain recommendations** (Alshammari, Nasraoui, & Sanders, 2019)**, semantic web based** (Alshammari, Nasraoui, & Sanders, 2019)**, SemAuto** (Alshammari, Nasraoui, & Sanders, 2019)**, SemRec** (Alshammari, Nasraoui, & Sanders, 2019)**, sensitivity analysis** (Arras, Horn, Montavon, Müller, & Samek, 2017) (Böhle, Eitel, Weygandt, & Ritter, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Roscher, Bohn, Duarte, & Garcke, 2020) (Rosenfeld & Richardson, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020)**, sensitivity analysis maps** (Rosenfeld & Richardson, 2019)**, sensitivity to local variation of the input image** (Arras, Horn, Montavon, Müller, & Samek, 2017)**, sentiment-based explainable recommendation** (Hou, Yang, Wu, & Yu, 2019)**, sentiment-based tradeoff-oriented explanation approach** (Chen, Yan, & Wang, 2019)**, separate engine for generating explanations in recommender systems** (Bharadhwaj & Joshi, 2018)**, shallow models** (Gunning & Aha, 2019)**, SHapley Additive exPlanations (SHAP)** (Ariza, Arroyo, Caparrini, & Segovia, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Pintelas, Livieris, & Pintelas, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020)**, Shapley Additive exPlanations (SHAP) values** (Ariza, Arroyo, Caparrini, & Segovia, 2020) (Zihni, et al., 2020)**, Shapley value sampling** (Singh, Sengupta, & Lakshminarayanan, 2020)**, Shapley Values** (Carvalho, Pereira, & Cardoso, 2019) (Käde & Von Maltzan, 2019) (Zihni, et al., 2020)**, shared tradeoff properties of a group of products in terms of both static specifications and feature sentiments** (Chen, Yan, & Wang, 2019)**, shared tradeoff properties of a group of products relative to the top recommendation** (Chen, Yan, & Wang, 2019)**, show the dataflow through the computational graph** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, Show-and-Tell Explanations** (Gunning & Aha, 2019)**, similar images** (Singh, Sengupta, & Lakshminarayanan, 2020)**, similarity analysis techniques** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, Simplified Tree Ensamble Learner (STEL)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Single Tree** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Single Tree Approximation (STA)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, SmoothGrad** (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Zihni, et al., 2020)**, SmoothGrad saliency maps** (Carvalho, Pereira, & Cardoso, 2019)**, social collaborative viewpoint regression** (Hong, Akerkar, & Jung, 2019)**, social explanations for**

recommender systems (Hong, Akerkar, & Jung, 2019)**, SP-LIME** (Käde & Von Maltzan, 2019) (Rosenfeld & Richardson, 2019)**, statistical feature importances** (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019)**, structured knowledge bases** (Alshammari, Nasraoui, & Sanders, 2019)**, surrogate model** (Carvalho, Pereira, & Cardoso, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, surrogates** (Roscher, Bohn, Duarte, & Garcke, 2020)**, susceptibility maps** (Böhle, Eitel, Weygandt, & Ritter, 2019)**, SVM margin** (Singh, Sengupta, & Lakshminarayanan, 2020)**, SVM+Prototypes (SVM+P)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, tag-based explaining approach in graph-based recommender** (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019)**, tags and ratings in a social tagging system with PARAFAC** (Hong, Akerkar, & Jung, 2019)**, Tagsplanations** (Hong, Akerkar, & Jung, 2019)**, TasteWeights** (Alshammari, Nasraoui, & Sanders, 2019)**, TCAV extension by Regression Concept Vectors (RCV)** (Singh, Sengupta, & Lakshminarayanan, 2020)**, TCAV extension by Uniform unit Ball surface Sampling (UBS)** (Singh, Sengupta, & Lakshminarayanan, 2020)**, TCAV with RCV** (Singh, Sengupta, & Lakshminarayanan, 2020)**, t-Distributed Stochastic Neighbor Embedding (t-SNE)** (Singh, Sengupta, & Lakshminarayanan, 2020)**, TempEx-Dry** (Bharadhwaj & Joshi, 2018)**, TempEx-Fluid** (Bharadhwaj & Joshi, 2018)**, Testing Concept Activation Vectors (TCAV)** (Singh, Sengupta, & Lakshminarayanan, 2020)**, text descriptions for pictures** (Rosenfeld & Richardson, 2019)**, text justifications** (Singh, Sengupta, & Lakshminarayanan, 2020)**, textual explanations** (Hong, Akerkar, & Jung, 2019) (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019) (Rosenfeld & Richardson, 2019)**, textual templates for pre-defined explanations** (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019)**, topic-based explainable recommendation** (Hou, Yang, Wu, & Yu, 2019)**, tractable probabilistic logic models (TPLMs)** (Gunning & Aha, 2019)**, tradeoff-oriented explanations** (Chen, Yan, & Wang, 2019)**, transparent models** (Rosenfeld & Richardson, 2019)**, Tree Metrics** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, tree regularization** (Carvalho, Pereira, & Cardoso, 2019)**, Tree Space Prototype (TSP)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, TreeView** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Trepan** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, tripartite graph encoding user-item-aspect relationships for a review-aware recommendation** (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019)**, Triplet loss** (Singh, Sengupta, & Lakshminarayanan, 2020)**, triplet-loss and k nearest neighbors (kNN) search-based learning strategy** (Singh, Sengupta, & Lakshminarayanan, 2020)**, TriRank** (Hong, Akerkar, & Jung, 2019)**, t-SNE (t-Distributed Stochastic Neighbor Embedding)** (Carvalho, Pereira, & Cardoso, 2019)**, two different networks to visualize predictions of different network layers** (Käde & Von Maltzan, 2019)**, Two-Level Boolean Rules (TLBR)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, UBS** (Singh, Sengupta, & Lakshminarayanan, 2020)**, U-Net based**

architecture and keypoints (Singh, Sengupta, & Lakshminarayanan, 2020)**, U-Net with shape attention stream** (Singh, Sengupta, & Lakshminarayanan, 2020)**, uniform probabilistic framework** (Preece, 2018)**, user-item relevance scores using matrix factorization techniques** (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019)**, users' sentiments on specific aspects** (Chen, Yan, & Wang, 2019)**, users' sentiments on specific features** (Chen, Yan, & Wang, 2019)**, Variable Effect Characteristic curve (VEC)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Variable Interaction Network (VIN)** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, Variable Interaction Network (VIN) visualization** (Rosenfeld & Richardson, 2019)**, vectors for localized interpretations** (Rosenfeld & Richardson, 2019)**, visible NNs** (Roscher, Bohn, Duarte, & Garcke, 2020)**, visual approaches** (Roscher, Bohn, Duarte, & Garcke, 2020)**, visual comparative analysis** (De Koninck, De Weerdt, & vanden Broucke, 2017)**, visual explanations** (Carvalho, Pereira, & Cardoso, 2019) (Hong, Akerkar, & Jung, 2019) (Rosenfeld & Richardson, 2019)**, visual heatmaps** (Singh, Sengupta, & Lakshminarayanan, 2020)**, visual word constraint** (Singh, Sengupta, & Lakshminarayanan, 2020)**, visualization** (Gunning & Aha, 2019) (Käde & Von Maltzan, 2019) (Singh, Sengupta, & Lakshminarayanan, 2020)**, visualization method** (Böhle, Eitel, Weygandt, & Ritter, 2019)**, visualization tools** (Rosenfeld & Richardson, 2019)**, visualizations** (Du, Liu, & Hu, 2019) (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019) (Preece, 2018) (Rosenfeld & Richardson, 2019)**, visualize convolutional filters** (Spinner, Schlegel, Schafer, & El-Assady, 2020)**, visualize filters and activations** (Böhle, Eitel, Weygandt, & Ritter, 2019)**, visualize the activations of each layer of a trained CNN** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, visualize the decision boundary in a two-dimensional plane** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, visualize the discrimination of data cohorts by means of projections guided by paths through the data (tours)** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, visualize the effect of individual inputs to the output** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)**, visualize the features of the different layers by regularized optimization in image space** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)**, visualize what computations and neuron activations occur in the intermediate layers of deep neural networks** (Rosenfeld & Richardson, 2019)**, What If Tool (WIT)** (Käde & Von Maltzan, 2019)**, XRL Interaction** (Gunning & Aha, 2019)**, z-LRP** (Spinner, Schlegel, Schafer, & El-Assady, 2020)

Appendix K. Explainability techniques terminology control

1Rule, 3-Level Explanation, Acceptance Testing, Accumulated Local Effects Plot, activation maps, AIF360, algorithmic transparency, analysis of layers of a 3D-CNN using Gaussian

mixture model (GMM) and binary encoding of training and test images based on their GMM components for returning similar 3D images, anchors, approximate an interpretable model for the black-box model, Argumentation and Pedagogy, argumentation theory based, Aspect-based Matrix Factorization model (AMF), attention heatmaps, attention maps, attention mask weights, attention-based model (attention based model, attention mechanism, attention modules)**, AttentiveChrome NN, attribute explanations** (attribute explanations, attribute identity explanations, attribute use explanations)**, attributions** (attribution maps, attributions)**, auditing, auxiliary criteria, auxiliary data, auxiliary network, backpropagation of the gradients, back-propagation-based methods** (back-propagation-based methods, backward propagation methods)**, Bayesian Case Model (BCM), Bayesian Rule Lists (BRL), Bayesian Teaching, BreakDown, CAM, capsule network, case-based reasoning** (case-based explanations, case-based reasoning (case-based explanations by example))**, causal explanations, causal models to explain learning (CAMEL) approach, CAV., CENTAUR, CFS, class activation mapping, class maps, CLEAR (CLass-Enhanced Attentive Response), clustering methods, coalitional game theory based** (coalition game theory to evaluate the effect of combinations of features, coalitional game theory based)**, co-clustering approach to gain explainability in a user-item bipartite network, collaborative filtering method using a tensor modelled by considering the 5Ws with explanations based on template, collaborative-based explanations, colour based nomogram, combination of genetic algorithms with decision trees or rules, combination with other machine learning methods, combine model compression with dimension reduction, combine physics-based models with machine learning, Combined Multiple Model (CCM), community tags to explain recommendations, complementary examples** (complemental examples, complementary examples)**, Compound Critiques, Concept Activation Vectors (TCAV)** (Concept Activation Vectors (TCAV), Testing Concept Activation Vectors (TCAV))**, concept vectors, concept-based explanations, conceptual clustering, Confident Decision Tree (CDT), Conj Rules, constraint programming for converting linear SVM (and other hyperplane-based linear classifiers) into a set of non overlapping and interpretable rules, content-based explanations, contextual decomposition explanation penalization, counterfactual explanations, CPAR (Classification based on Predictive Association Rules), cross-cluster model-log alignment for identifying differences between clusters, CTR, data points, data-dependent, dataset-level, Dead Weight, Decision Diagrams , decision rules, decision rules, decision trees, decompositional rule extraction** (decomposability, decompositional rule extraction)**, DeConvNet, deconvolution** (deconvolution, deconvolutional networks)**, DecText, Deep Attention-Based Representations for Explanation/Explainable Generative Adversarial Networks (DARE/X-GANS), deep explanation, Deep Hierarchical Generative Models, Deep Learning Important FeaTures (DeepLIFT) , Deep SHapley Additive exPlanations, deep Taylor decomposition, deep tensor networks, DeepDreams, DeepExplain, DeepTaylor, DeepTune, descriptive statistics, design**

transparency, diagnostic sentence, domain knowledge supported (domain constraints, domain knowledge supported interpretation, domain-based feature engineering, domain-dependent)**, dual neural network system, dual system approach, EG, e-LRP, enforce sparsity terms, ERBM, example-based explanation** (example-based explanation, examples)**, excitable network attractors (ENAs), expert knowledge, expert-determined features relevance, explain the recommendation process, Explainable Expert Systems (EES) project, Explainable Matrix Factorization (EMF), explainable question answering system (EQUAS), explainable recommendation, explainer, explanation in terms of input variables, explanation maps, explanation-based learning, explanations for a hybrid recommender system, explanations in time-series recommendation, explicit factor model, explicitly capturing and displaying the interactions learned by a neural network, Factorized Latent Aspect ModEl (FLAME) combining collaborative filtering and opinion mining, feature analysis** (feature analysis, feature importance analysis)**, feature attribution , feature engineering, feature extraction and explanation extraction framework, feature importance** (feature importance, feature importance analysis, feature importance and ranking, feature importance scores, permutation feature importance, relevance of input pixels, relevance scores)**, feature influence** (feature influence methods, input variables influence)**, Feature Interaction, feature sentiments, feature summary** (feature summary, feature summary statistic, feature summary visualization, statistical feature importances)**, filters, FINE (feature importance in nonlinear embeddings), Forest Floor, four-order tensor to model users, FRL, Gaussian Process Classification (GDP), Generalized Additive Models (GAMs), generative modelling, genetic programming, global explanation** (global, global explanation methods, global in scope, global interpretability, global level of explainability, global model interpretability)**, GMM and atlas, GNNExplainer, GoldenEye, GPDT, grad*input** (grad*input, gradient x input, Gradient*I/P, Input × Gradient)**, GradHM+AS, GradHM+TS , Gradient weighted class activation mapping (GradCAM)** (Gradient weighted class activation mapping (GradCAM), Grad-CAM)**, gradient-based** (gradient, gradient methods, gradient-based approaches, gradient-based methods)**, gradient-weighted heatmap (GradHM), graph-based recommendation approach, graphing the functional relationship between the predicted response and the feature for individual observations, G-REX, GroupINN, GSInquire, guided backpropagation, Guided Grad-CAM, Guided-GradHM, Guided-GradHM+AS, Guided-GradHM+TS, heatmaps** (visual heatmaps, heat maps, heatmaps, heatmaps of salient regions)**, HFT, HIN technique, histogram, HistoTrend, hybrid approach using collaborative and content-based filtering techniques, image reconstruction, individual conditional expectation, inductive logic programming, Influence Functions, influential data points, Info Flow, Information Plane** (Information Plane, Information Plane visualization)**, in-model, in-model joint architecture from explainer and classifier, iNNvestigate, instance-level explanations, integrate explanations into Matrix Factorization, integrated gradients** (Integrated Grad, integrated gradients)**, integration of**

feature interaction and tree interpretation functionalities into Random Forest program code, interaction and feature importance, Interactive Training, Interactive Visualization, interpretability constraints (interpretability constraints, interpretability constraints into the structure of the model)**, interpretable convolutional neural networks, Interpretable Decision Sets (IDS), interpretable model extraction, interpretable models, InterpretML-Framework, inTrees, intrinsic explainability** (intrinsic, intrinsic explainability, intrinsic explanation, intrinsic interpretability, intrinsic methods)**, intrinsic interpretable Grey-Box ensemble model, Introspective explanations, investigation of deep representations, items, ITLFM, justification explanations;, justify why the recommendation might be good for a user, keywords and neighbours and ratings, keywords or user-tags based explanations, k-means, kNN, knowledge distillation, knowledge extraction, knowledge-based explanations, Layer wise relevance propagation (LRP), leverage topic models to discover explainable latent factors in matrix factorization, linear dimensionality reduction, linear models, linked data, local approximation-based explanation, local explanation** (local, local explanation, local in scope, local interpretability, local interpretations, local level of explainability, local model interpretability, local-level)**, Local Interpretable Model-Agnostic Explanation (LIME), Local Surrogate Model, LSTM, Mapping between image to reports, mask perturbation, maximum mean discrepancy** (MMD-critic (Maximum Mean Discrepancy), maximum mean discrepancy)**, MDNet, mimic learning** (interpretable mimic learning, mimic learning)**, MinMax, MMALFM, model coefficients for logistic regression, model compression, Model Explanation System (MES), model explanations, model induction, model internals, model modularity, model simulatability, model sparsity, model tool, model-agnostic explanations** (model agnostic explanations, model-agnostic, agnostic explanators)**, model-based, model-based feature engineering, model-specific, monotonic constraints, MoviExplain, multivariate filters, MYCIN, Narrative Generation, natural language caption generation, nearest neighbours, Neighbour ratings, neighbourhood of an instance, neighbourhood technique based on cosine similarity, neighbourhood based Collaborative Filtering (CF), neighbourhood style explanation, NEOMYCIN, network propagation technique based on deconvolutions to reconstruct input image patterns that are linked to a particular feature map activation or prediction, neural activation visualization, neural interaction detection, Neural Interpretation Diagram (NID), neural rating and tips generation, Neurons Activation (NA), NL explanations** (NL explanations, NL generation, NL justifications, text justifications, textual explanations)**, Node-Link Vis, nonlinear dimensionality reduction, oblique tree sparse additive models (OT-SpAMs), occlusion, One-variable-at-a-Time approach, optimal selection of teaching examples, Orthogonal Projection of Input Attributes (OPIA), outcome tool, PALM, Partial Dependency Plots (PDP)** (partial dependence bars, Partial Dependence Plot, Partial Dependency Plots (PDP))**, Pattern Attribution, PatternNet, pedagogical rule extraction, perturbation-based explanation**

(perturbation, perturbation-based explanation, random perturbations), post hoc application of supervised learning with support vector machines, post-hoc explanation (post hoc interpretability, post-hoc, post-hoc analysis, post-hoc explanation, post-hoc interpretability, post-hoc methods, post-hoc rationalisation, post-model)**, post-hoc mechanism to generate explanations in recommender systems, prediction-level, Preference-based Organization (Pref-ORG), pre-model, Principal Component Analysis, probabilistic generative model, Prospector, prototype and criticism generation** (prototype and criticism generation, prototypes and criticisms)**, prototype generation by greedy approach, prototype generation by LP relaxation with randomized rounding, Prototype Selection (PS), prototypes** (prototype, identification of prototypes, identify prototypes, prototype analysis)**, proxy models, Quantitative Input Influence (QII), Query evidence that explains DNN decisions, random sampling based, Rational explanations, rationales as part of the learning process, RBLT, reducing complex NN, Reflexive explanations, REFNE, relational connecting paths, RETAIN (REverse Time AttentIoN), reverse engineering, reviews with ratings to enhance the explainability of matrix factorization, RippleNet, RMR, root causes of misclassifications, root causes of process model differences, Rule Based Explanator, rule extraction, Rule Set, rule-based methods, rule-based methods for recommender systems , rule-based segmentation , rule-based segmentation followed by a perturbation analysis , RxREN, saliency, saliency heatmaps** (saliency heatmaps, heatmaps of salient regions)**, saliency maps, saliency masks, salient (highest weighted or most predictive) text features or fragments, salient examples, salient part of the images, salient sentences from text documents using loss gradient magnitudes, salient structures within images related to a specific class by computing the corresponding prediction score derivative with respect to the input image, Saturated Weight, SAUNet, search for explanations of clusters of process instances (SECPI), second deep network that generates explanations, self-organizing maps, self-training Grey-Box model, semantic distance (LDSD) algorithm and DBpedia based recommendation system, semantic meaningfulness constraints, semantic monotonicity constraints, semantic property values to explain recommendations, semantic web based, SemAuto, SemRec, sensitivity analysis, sensitivity analysis maps, sensitivity to local variation of the input image, sentiment-based explainable recommendation, sentiment-based tradeoff-oriented explanation approach, separate engine for generating explanations in recommender systems, shallow models, Shapley Additive exPlanations (SHAP) values** (SHapley Additive exPlanations (SHAP), Shapley Additive exPlanations (SHAP) values, Shapley value sampling, Shapley Values)**, shared trade-off properties of a group of products in terms of both static specifications and feature sentiments, shared trade-off properties of a group of products relative to the top recommendation, show the dataflow through the computational graph, Show-and-Tell Explanations, similar images, similarity analysis techniques, Simplified Tree Ensamble Learner (STEL), Single Tree, Single Tree**

Approximation (STA), SmoothGrad, SmoothGrad saliency maps, social collaborative viewpoint regression, social explanations for recommender systems, SP-LIME, structured knowledge bases, surrogate model (surrogate model, surrogates)**, susceptibility maps, SVM margin, SVM+Prototypes (SVM+P), tag-based explaining approach in graph-based recommender, tags and ratings in a social tagging system with PARAFAC, Tagsplanations, TasteWeights, TCAV extension by Regression Concept Vectors (RCV)** (TCAV extension by Regression Concept Vectors (RCV), TCAV with RCV)**, TCAV extension by Uniform unit Ball surface Sampling (UBS), t-Distributed Stochastic Neighbor Embedding (t-SNE), TempEx-Dry, TempEx-Fluid, text descriptions for pictures, textual templates for pre-defined explanations, topic-based explainable recommendation, tractable probabilistic logic models (TPLMs), trade-off-oriented explanations, transparent models, Tree Metrics, tree regularization, Tree Space Prototype (TSP), TreeView, Trepan, tripartite graph encoding user-item-aspect relationships for a review-aware recommendation, Triplet loss, triplet-loss and k nearest neighbours (kNN) search-based learning strategy, TriRank, t-SNE (t-Distributed Stochastic Neighbour Embedding), two different networks to visualize predictions of different network layers, Two-Level Boolean Rules (TLBR), UBS, U-Net based architecture and key points, U-Net with shape attention stream, uniform probabilistic framework, user-item relevance scores using matrix factorization techniques, users' sentiments on specific aspects, users' sentiments on specific features, Variable Effect Characteristic curve (VEC), Variable Interaction Network (VIN)** (Variable Interaction Network (VIN), Variable Interaction Network (VIN) visualization)**, vectors for localized interpretations, visible NNs, visual comparative analysis, visual word constraint, visualizations** (visual approaches, visual explanations, visualization, data visualization methods, visualization method, visualization tools, visualizations)**, visualize convolutional filters, visualize filters and activations, visualize the activations of each layer of a trained CNN, visualize the decision boundary in a two-dimensional plane, visualize the discrimination of data cohorts by means of projections guided by paths through the data (tours), visualize the effect of individual inputs to the output, visualize the features of the different layers by regularized optimization in image space, visualize what computations and neuron activations occur in the intermediate layers of deep neural networks, What If Tool (WIT), XRL Interaction, z-LRP.**

Appendix L. Explainability techniques

1Rule, 3-Level Explanation, Acceptance Testing, Accumulated Local Effects Plot, activation maps, AIF360, algorithmic transparency, analysis of layers of a 3D-CNN using Gaussian mixture model (GMM) and binary encoding of training and test images based on their GMM components

for returning similar 3D images, anchors, approximate an interpretable model for the black-box model, Argumentation and Pedagogy, argumentation theory based, Aspect-based Matrix Factorization model (AMF), attention heatmaps, attention maps, attention mask weights, attention-based model, AttentiveChrome NN, attribute explanations, attributions, auditing, auxiliary criteria, auxiliary data, auxiliary network, backpropagation of the gradients, back-propagation-based methods, Bayesian Case Model (BCM), Bayesian Rule Lists (BRL), Bayesian Teaching, BreakDown, CAM, capsule network, case-based reasoning, causal explanations, causal models to explain learning (CAMEL) approach, CAV., CENTAUR, CFS, class activation mapping, class maps, CLEAR (CLass-Enhanced Attentive Response), clustering methods, coalitional game theory based, co-clustering approach to gain explainability in a user-item bipartite network, collaborative filtering method using a tensor modeled by considering the 5Ws with explanations based on template, collaborative-based explanations, color based nomogram, combination of genetic algorithms with decision trees or rules, combination with other machine learning methods, combine model compression with dimension reduction, combine physics-based models with machine learning, Combined Multiple Model (CCM), community tags to explain recommendations, complementary examples, Compound Critiques, Concept Activation Vectors (TCAV), concept vectors, concept-based explanations, conceptual clustering, Confident Decision Tree (CDT), Conj Rules, constraint programming for converting linear SVM (and other hyperplane-based linear classifiers) into a set of non overlapping and interpretable rules, content-based explanations, contextual decomposition explanation penalization, counterfactual explanations, CPAR (Classification based on Predictive Association Rules), cross-cluster model-log alignment for identifying differences between clusters, CTR, data points, data-dependent, dataset-level, Dead Weight, Decision Diagrams, decision rules, decision rules, decision trees, decompositional rule extraction, DeConvNet, deconvolution, DecText, Deep Attention-Based Representations for Explanation/Explainable Generative Adversarial Networks (DARE/X-GANS), deep explanation, Deep Hierarchical Generative Models, Deep Learning Important FeaTures (DeepLIFT) , Deep SHapley Additive exPlanations, deep Taylor decomposition, deep tensor networks, DeepDreams, DeepExplain, DeepTaylor, DeepTune, descriptive statistics, design transparency, diagnostic sentence, domain knowledge supported, dual neural network system, dual system approach, EG, e-LRP, enforce sparsity terms, ERBM, example-based explanation, excitable network attractors (ENAs), expert knowledge, expert-determined features relevance, explain the recommendation process, Explainable Expert Systems (EES) project, Explainable Matrix Factorization (EMF), explainable question answering system (EQUAS), explainable recommendation, explainer, explanation in terms of input variables, explanation maps, explanation-based learning, explanations for a hybrid recommender system, explanations in time-series recommendation, explicit factor model, explicitly capturing and displaying the interactions learned by a neural network, Factorized Latent Aspect ModEl (FLAME) combining collaborative filtering and opinion mining, feature analysis, feature attribution, feature

*engineering, feature extraction and explanation extraction framework, feature importance, feature influence, Feature Interaction, feature sentiments, feature summary, filters, FINE (feature importance in nonlinear embeddings), Forest Floor, four-order tensor to model users, FRL, Gaussian Process Classification (GDP), Generalized Additive Models (GAMs), generative modeling, genetic programming, global explanation, GMM and atlas, GNNExplainer, GoldenEye, GPDT, grad*input, GradHM+AS, GradHM+TS, Gradient weighted class activation mapping (GradCAM), gradient-based, gradient-weighted heatmap (GradHM), graph-based recommendation approach, graphing the functional relationship between the predicted response and the feature for individual observations, G-REX, GroupINN, GSInquire, guided backpropagation, Guided Grad-CAM, Guided-GradHM, Guided-GradHM+AS, Guided-GradHM+TS, heatmaps, HFT, HIN technique, histogram, HistoTrend, hybrid approach using collaborative and content-based filtering techniques, image reconstruction, individual conditional expectation, inductive logic programming, Influence Functions, influential data points, Info Flow, Information Plane, in-model, in-model joint architecture from explainer and classifier, iNNvestigate, instance-level explanations, integrate explanations into Matrix Factorization, integrated gradients, integration of feature interaction and tree interpretation functionalities into Random Forest program code, interaction and feature importances, Interactive Training, Interactive Visualization, interpretability constraints, interpretable convolutional neural networks, Interpretable Decision Sets (IDS), interpretable model extraction, interpretable models, InterpretML-Framework, inTrees, intrinsic explainability, intrinsic interpretable Grey-Box ensemble model, Introspective explanations, investigation of deep representations, items, ITLFM, justification explanations;, justify why the recommendation might be good for a user, keywords and neighbours and ratings, keywords or user-tags based explanations, k-means, kNN, knowledge distillation, knowledge extraction, knowledge-based explanations, Layer wise relevance propagation (LRP), leverage topic models to discover explainable latent factors in matrix factorization, linear dimensionality reduction, linear models, linked data, local approximation-based explanation, local explanation, Local Interpretable Model-Agnostic Explanation (LIME), Local Surrogate Model, LSTM, Mapping between image to reports, mask perturbation, maximum mean discrepancy, MDNet, mimic learning, MinMax, MMALFM, model coefficients for logistic regression, model compression, Model Explanation System (MES), model explanations, model induction, model internals, model modularity, model simulatability, model sparsity, model tool, model-agnostic explanations, model-based, model-based feature engineering, model-specific, monotonic constraints, MoviExplain, multivariate filters, MYCIN, Narrative Generation, natural language caption generation, nearest neighbors, Neighbor ratings, neighborhood of an instance, neighborhood technique based on cosine similarity, neighbourhood based Collaborative Filtering (CF), neighbourhood style explanation, NEOMYCIN, network propagation technique based on deconvolutions to reconstruct input image patterns that are linked to a particular feature map activation or prediction, neural activation*

visualization, neural interaction detection, Neural Interpretation Diagram (NID), neural rating and tips generation, Neurons Activation (NA), NL explanations, Node-Link Vis, nonlinear dimensionality reduction, oblique tree sparse additive models (OT-SpAMs), occlusion, One-variable-at-a-Time approach, optimal selection of teaching examples, Orthogonal Projection of Input Attributes (OPIA), outcome tool, PALM, Partial Dependency Plots (PDP), Pattern Attribution, PatternNet, pedagogical rule extraction, perturbation-based explanation (perturbation, perturbation-based explanation, random perturbations), post hoc application of supervised learning with support vector machines, post-hoc explanation, post-hoc mechanism to generate explanations in recommender systems, prediction-level, Preference-based Organization (Pref-ORG), pre-model, Principal Component Analysis, probabilistic generative model, Prospector, prototype and criticism generation, prototype generation by greedy approach, prototype generation by LP relaxation with randomized rounding, Prototype Selection (PS), prototypes, proxy models, Quantitative Input Influence (QII), Query evidence that explains DNN decisions, random sampling based, Rational explanations, rationales as part of the learning process, RBLT, reducing complex NN, Reflexive explanations, REFNE, relational connecting paths, RETAIN (REverse Time AttentIoN), reverse engineering, reviews with ratings to enhance the explainability of matrix factorization, RippleNet, RMR, root causes of misclassifications, root causes of process model differences, Rule Based Explanator, rule extraction, Rule Set, rule-based methods, rule-based methods for recommender systems, rule-based segmentation, rule-based segmentation followed by a perturbation analysis, RxREN, saliency, saliency heatmaps, saliency maps, saliency masks, salient (highest weighted or most predictive) text features or fragments, salient examples, salient part of the images, salient sentences from text documents using loss gradient magnitudes, salient structures within images related to a specific class by computing the corresponding prediction score derivative with respect to the input image, Saturated Weight, SAUNet, search for explanations of clusters of process instances (SECPI), second deep network that generates explanations, self-organizing maps, self-training Grey-Box model, semantic distance (LDSD) algorithm and DBpedia based recommendation system, semantic meaningfulness constraints, semantic monotonicity constraints, semantic property values to explain recommendations, semantic web based, SemAuto, SemRec, sensitivity analysis, sensitivity analysis maps, sensitivity to local variation of the input image, sentiment-based explainable recommendation, sentiment-based tradeoff-oriented explanation approach, separate engine for generating explanations in recommender systems, shallow models, Shapley Additive exPlanations (SHAP) values, shared tradeoff properties of a group of products in terms of both static specifications and feature sentiments, shared tradeoff properties of a group of products relative to the top recommendation, show the dataflow through the computational graph, Show-and-Tell Explanations, similar images, similarity analysis techniques, Simplified Tree Ensamble Learner (STEL), Single Tree, Single Tree Approximation (STA), SmoothGrad, SmoothGrad saliency maps, social collaborative viewpoint regression, social explanations for recommender systems, SP-

LIME, structured knowledge bases, surrogate model, susceptibility maps, SVM margin, SVM+Prototypes (SVM+P), tag-based explaining approach in graph-based recommender, tags and ratings in a social tagging system with PARAFAC, Tagsplanations, TasteWeights, TCAV extension by Regression Concept Vectors (RCV), TCAV extension by Uniform unit Ball surface Sampling (UBS), t-Distributed Stochastic Neighbor Embedding (t-SNE), TempEx-Dry, TempEx-Fluid, text descriptions for pictures, textual templates for pre-defined explanations, topic-based explainable recommendation, tractable probabilistic logic models (TPLMs), tradeoff-oriented explanations, transparent models, Tree Metrics, tree regularization, Tree Space Prototype (TSP), TreeView, Trepan, tripartite graph encoding user-item-aspect relationships for a review-aware recommendation, Triplet loss, triplet-loss and k nearest neighbors (kNN) search-based learning strategy, TriRank, t-SNE (t-Distributed Stochastic Neighbor Embedding), two different networks to visualize predictions of different network layers, Two-Level Boolean Rules (TLBR), UBS, U-Net based architecture and keypoints, U-Net with shape attention stream, uniform probabilistic framework, user-item relevance scores using matrix factorization techniques, users' sentiments on specific aspects, users' sentiments on specific features, Variable Effect Characteristic curve (VEC), Variable Interaction Network (VIN), vectors for localized interpretations, visible NNs, visual comparative analysis, visual word constraint, visualizations, visualize convolutional filters, visualize filters and activations, visualize the activations of each layer of a trained CNN, visualize the decision boundary in a two-dimensional plane, visualize the discrimination of data cohorts by means of projections guided by paths through the data (tours), visualize the effect of individual inputs to the output, visualize the features of the different layers by regularized optimization in image space, visualize what computations and neuron activations occur in the intermediate layers of deep neural networks, What If Tool (WIT), XRL Interaction, z-LRP.

Appendix M. Credit scoring model components to social media user profiling

Bank-borrower relationship

→ none

Collateral characteristics

→ **Artificial neural networks** (Xu, et al., 2019) (Xu, et al., 2019) (Zhang, et al., 2018)

→ **Clustering models** (Béjar, et al., 2016) (Xu, et al., 2019)

→ **Decision Trees** (Xu, et al., 2019)

→ **Ensemble Learning** (Liao, et al., 2015) (Xu, et al., 2019) (Zhang, et al., 2018) (Zhuang, et al., 2017)

→ **Graph Theory algorithms** (Al Hasan Haldar, et al., 2019) (Liao, et al., 2015) (Xu, et al., 2019) (Zhang, et al., 2018)

→ **Linear models** (Al Hasan Haldar, et al., 2019) (Xu, et al., 2019) (Zhang, et al., 2018)

→ **Natural Language Processing** (Al Hasan Haldar, et al., 2019) (Béjar, et al., 2016) (Xu, et al., 2014) (Ta, et al., 2019) (Xu, et al., 2019) (Zhang, et al., 2018) (Zhuang, et al., 2017)

→ **Probabilistic and statistical models** (Al Hasan Haldar, et al., 2019) (Béjar, et al., 2016) (Xu, et al., 2014) (Liao, et al., 2015) (Ta, et al., 2019) (Xu, et al., 2019) (Xu, et al., 2019) (Zhuang, et al., 2017)

→ **Social Recommender Systems** (Al Hasan Haldar, et al., 2019) (Ta, et al., 2019) (Zhang, et al., 2018)

→ **Support Vector Machines** (Xu, et al., 2019) (Zhang, et al., 2018) (Zhuang, et al., 2017)

Credit applied for

→ none

Credit card(s) data

→ none

Credit history

→ none

Demographic data

→ **Artificial Neural Networks** (Li, et al., 2019) (Li, et al., 2019) (Xu, et al., 2019) (Xu, et al., 2019) (Zhang, et al., 2018) (Zheng, et al., 2019)

→ **Clustering models** (De Salve, et al., 2018) (Dougnon, et al., 2015) (Dougnon, et al., 2016) (Eke, et al., 2019) (Li, et al., 2019) (Xu, et al., 2019)

→ **Decision trees** (Barbon, et al., 2017) (Dougnon, et al., 2016) (Gu, et al., 2018) (Guo, et al., 2016) (Peng, et al., 2017) (Xu, et al., 2019)

→ **Ensemble learning** (Barbon, et al., 2017) (Dougnon, et al., 2015) (Dougnon, et al., 2016) (Gu, et al., 2018) (Guo, et al., 2016) (Hirt, et al., 2019) (Li, et al., 2019) (Liao, et al., 2015) (Xu, et al., 2019) (Zhang, et al., 2018) (Zheng, et al., 2019)

→ **Graph theory algorithms** (Barbon, et al., 2017) (Dougnon, et al., 2015) (Dougnon, et al., 2016) (Liao, et al., 2015) (Peng, et al., 2017) (Xu, et al., 2019) (Zhang, et al., 2018)

→ **Linear models** (Dougnon, et al., 2015) (Dougnon, et al., 2016) (Gu, et al., 2018) (Guo, et al., 2016) (Li, et al., 2019) (Xu, et al., 2019) (Zhang, et al., 2018) (Zheng, et al., 2019)

→ **Natural Language Processing** (Barbon, et al., 2017) (Dougnon, et al., 2015) (Dougnon, et al., 2016) (Hirt, et al., 2019) (Li, et al., 2019) (Wang, et al., 2017) (Xu, et al., 2019) (Zhang, et al., 2018)

→ **Nearest neighbor models** (Barbon, et al., 2017) (Li, et al., 2019) (Peng, et al., 2017)

→ **Probabilistic and statistical models** (Barbon, et al., 2017) (Dougnon, et al., 2015) (Dougnon, et al., 2016) (Gu, et al., 2018) (Guo, et al., 2016) (Li, et al., 2019) (Li, et al., 2019) (Liao, et al., 2015) (Peng, et al., 2017) (Fang, et al., 2015) (Xu, et al., 2019) (Xu, et al., 2019)

→ Social recommender systems (Zhang, et al., 2018)

→ **Social semantic web** (Dougnon, et al., 2016) (Li, et al., 2019) (Wang, et al., 2017)

→ **Support vector machines** (Barbon, et al., 2017) (Dougnon, et al., 2016) (Gu, et al., 2018) (Guo, et al., 2016) (Li, et al., 2019) (Li, et al., 2019) (Peng, et al., 2017) (Fang, et al., 2015) (Wang, et al., 2017) (Xu, et al., 2019) (Zhang, et al., 2018)(Zheng, et al., 2019)

Employment status

→ **Artificial Neural Networks** (Li, et al., 2019) (Tong, et al., 2016)

→ Dimensionality reduction (Tong, et al., 2016)

→ **Ensemble learning** (Guo, et al., 2016) (Huang, et al., 2015) (Liao, et al., 2015) (Tong, et al., 2016)

→ **Graph theory algorithms** (Liao, et al., 2015) (Tong, et al., 2016)

→ **Natural Language Processing** (Huang, et al., 2015) (Tong, et al., 2016) (Wang, et al., 2017)

→ **Nearest neighbour models** (Eke, et al., 2019) (Huang, et al., 2015) (Li, et al., 2019)

→ **Probabilistic and statistical models** (Guo, et al., 2016) (Huang, et al., 2015) (Li, et al., 2019) (Liao, et al., 2015) (Fang, et al., 2015)

→ Social semantic web (Wang, et al., 2017)

→ **Support vector machines** (Guo, et al., 2016) (Li, et al., 2019) (Fang, et al., 2015) (Wang, et al., 2017)

Financial indicators (income vs. expenses)

→ **Clustering models** (Lampos, et al., 2016)

→ **Decision trees** (Guo, et al., 2016)

→ Ensemble learning (Guo, et al., 2016)

→ **Linear models** (Guo, et al., 2016)

→ **Probabilistic and statistical models** (Guo, et al., 2016) (Lampos, et al., 2016)

→ Support vector machines (Guo, et al., 2016)

Look-a-likes

→ **Artificial Neural Networks** (C C, et al., 2019) (Chen, et al., 2018) (Chen, et al., 2018) (Eke, et al., 2019) (Huang, et al., 2017) (Kang, et al., 2019) (De Salve, et al., 2018) (Ma, et al., 2019) (Pla Karidi, et al., 2018) (Wang, et al., 2018) (Xie, et al., 2018) (You, et al., 2016) (Zhang, et al., 2019)

→ **Clustering models** (Al-Qurishi, et al., 2018) (Alshammari, et al., 2019) (Anand, et al., 2014) (Arain, et al., 2017) (Barysheva, et al., 2015) (C C, et al., 2019) (Chen, et al., 2014) (De Salve, et al., 2018) (Gorrab, et al., 2017) (Hoang, et al., 2017) (Huang, et al., 2017) (Ju, et al., 2017) (Kumar, et al., 2012) (Kumar, et al., 2012) (Laere, et al., 2014) (Laere, et al., 2017) (Logesh, et al., 2019) (Ma, et al., 2019) (Pang, et al., 2013) (Wang, et al., 2018) (Yin, et al., 2018) (You, et al., 2016)

→ **Decision trees** (Chen, et al., 2014) (Faralli, et al., 2015) (Lee, et al., 2018) (Tang, et al., 2010) (Xie, et al., 2018) (Zhang, et al., 2019)

→ **Ensemble learning** (Chen, et al., 2018) (França, et al., 2018) (Liao, et al., 2015) (Zhang, et al., 2019)

→ **Graph theory algorithms** (Al-Qurishi, et al., 2018) (Barysheva, et al., 2015) (Chen, et al., 2018) (Chen, et al., 2018) (Chen, et al., 2014) (Kumar, et al., 2012) (Kumar, et al., 2012) (Lee, et al., 2018) (Liao, et al., 2015) (Pereira, et al., 2018) (Pla Karidi, et al., 2018) (Sultana, et al., 2016) (Yin, et al., 2018) (Zarrinkalam, et al., 2019)

→ **Linear models** (Chen, et al., 2014) (Pang, et al., 2013) (Pipanmaekaporn, et al., 2015) (Zhang, et al., 2019)

→ **Natural Language Processing** (Al-Qurishi, et al., 2018) (Anand, et al., 2014) (Anand, et al., 2014) (Arain, et al., 2017) (Barysheva, et al., 2015) (Bennacer Seghouani, et al., 2019) (Chen, et al., 2018) (Chen, et al., 2017) (França, et al., 2018) (Gorrab, et al., 2017) (Hoang, et al., 2017) (Huang, et al., 2017) (Kang, et al., 2019) (Ma, et al., 2015) (Nicoletti, et al., 2013) (Pang, et al., 2013) (Pereira, et al., 2018) (Piao, et al., 2018) (Pipanmaekaporn, et al., 2015) (Pla Karidi, et al., 2018) (Wu, et al., 2016) (Xie, et al., 2018) (Zarrinkalam, et al., 2019) (Zhang, et al., 2019) (Zheng, et al., 2018)

→ **Nearest neighbour models** (Anand, et al., 2014) (Arain, et al., 2017) (C C, et al., 2019) (Eke, et al., 2019) (Pang, et al., 2013) (Piao, et al., 2018) (Pla Karidi, et al., 2018) (Syed Mustapha, 2018) (Valsamis, et al., 2017)

→ **Probabilistic and statistical models** (Arain, et al., 2017) (Bennacer Seghouani, et al., 2019) (C C, et al., 2019) (Chen, et al., 2018) (Chen, et al., 2017) (Chen, et al., 2014) (Eyharabide, et al., 2012) (Hoang, et al., 2017) (Huang, et al., 2017) (Jansen, et al., 2018) (Kumar, et al., 2012) (Kumar, et al., 2012) (Laere, et al., 2014) (Laere, et al., 2017) (Lee, et al., 2018) (Liao, et al., 2015) (Ma, et al., 2015) (Ma, et al., 2019) (Manca, et al., 2018) (Pang, et al., 2013) (Piao, et al., 2018) (Pipanmaekaporn, et al., 2015) (Pla Karidi, et al., 2018) (Tang, et al., 2010) (Wang, et al., 2018) (Wu, et al., 2016) (Xie, et al., 2018) (Zarrinkalam, et al., 2019) (Zhang, et al., 2019) (Zheng, et al., 2018) (Zhou, et al., 2015) (Zhou, et al., 2012)

→ **Social recommender systems** (Al-Qurishi, et al., 2018) (Anand, et al., 2014) (Anand, et al., 2014) (Arain, et al., 2017) (C C, et al., 2019) (Chen, et al., 2017) (Chen, et al., 2018) (Dharia, et al., 2018) (Faralli, et al., 2015) (Jansen, et al., 2018) (Ju, et al., 2017) (Logesh, et al., 2019) (Ma, et al., 2019) (Pla Karidi, et al., 2018) (Tang, et al., 2014) (Valsamis, et al., 2017) (Wang, et al., 2018) (Wu, et al., 2016) (Xie, et al., 2018) (Yang, et al., 2015) (Zarrinkalam, et al., 2019) (Zheng, et al., 2018) (Zhou, et al., 2012)

→ **Social semantic web** (Besel, et al., 2016) (C C, et al., 2019) (Eyharabide, et al., 2012) (Laere, et al., 2014) (Laere, et al., 2017) (Lee, et al., 2018) (Lully, et al., 2018) (Orlandi, 2012) (Peña, et al., 2013) (Piao, et al., 2018) (Pla Karidi, et al., 2018) (Syed Mustapha, 2018) (Tang, et al., 2010) (Xie, et al., 2018) (Yin, et al., 2018)

→ **Support vector machines** (Chen, et al., 2018) (Chen, et al., 2014) (Faralli, et al., 2015) (Hoang, et al., 2017) (Lee, et al., 2018) (Ma, et al., 2015) (Pang, et al., 2013) (Pipanmaekaporn, et al., 2015) (Tang, et al., 2010) (Zhang, et al., 2019)

Psychological variables

→ **Artificial Neural Networks** (Buraya, et al., 2018) (Eke, et al., 2019)

→ **Ensemble learning** (Buraya, et al., 2018)

→ Graph theory algorithms (Kandias, et al., 2014)

→ **Linear models** (Buraya, et al., 2018) (Kandias, et al., 2014)

→ Natural Language Processing (Buraya, et al., 2018)

→ **Probabilistic and statistical models** (Buraya, et al., 2018) (Kandias, et al., 2014) (Fang, et al., 2015)

→ Social recommender systems (Buraya, et al., 2018)

→ **Support vector machines** (Kandias, et al., 2014) (Fang, et al., 2015)

Semiometric space

→ **Artificial Neural Networks** (Buraya, et al., 2018) (Eke, et al., 2019)

→ **Ensemble learning** (Buraya, et al., 2018)

→ **Linear models** (Buraya, et al., 2018)

→ Natural Language Processing (Buraya, et al., 2018)

→ Probabilistic and statistical models (Buraya, et al., 2018)

→ Social recommender systems (Buraya, et al., 2018)

Social network data

→ **Artificial Neural Networks** (C C, et al., 2019) (Chen, et al., 2018) (Chen, et al., 2018) (Ma, et al., 2019) (Xie, et al., 2018)

→ **Clustering models** (C C, et al., 2019) (Chen, et al., 2014) (Logesh, et al., 2019) (Ma, et al., 2019)

→ **Graph theory algorithms** (Chen, et al., 2018) (Chen, et al., 2018) (Chen, et al., 2014) (Pereira, et al., 2018)

→ **Linear models** (Chen, et al., 2014)

→ **Natural Language Processing** (Chen, et al., 2018) (Ma, et al., 2015) (Pereira, et al., 2018) (Xie, et al., 2018)

→ **Nearest neighbour models** (C C, et al., 2019) (Eke, et al., 2019)

→ **Probabilistic and statistical models** (C C, et al., 2019) (Chen, et al., 2018) (Chen, et al., 2014) (Ma, et al., 2015) (Ma, et al., 2019) (Xie, et al., 2018)

→ **Social recommender systems** (C C, et al., 2019) (Chen, et al., 2018) (Logesh, et al., 2019) (Ma, et al., 2019) (Xie, et al., 2018)

→ **Social semantic web** (C C, et al., 2019) (Xie, et al., 2018)

User-generated content

→ none

Appendix N. Social media user profiling approaches to explainability techniques

Artificial Neural Networks

→ **Decision Tree based explanations** (Rosenfeld & Richardson, 2019) (Carvalho, Pereira, & Cardoso, 2019) (Arun, 2020) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Du, Liu, & Hu, 2019)

→ **Deep explanations** (Preece, 2018) (Carvalho, Pereira, & Cardoso, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020) (Böhle, Eitel, Weygandt, & Ritter, 2019) (Karlsson, Rebane, Papapetrou, & Gionis, 2019) (Arras, Horn, Montavon, Müller, & Samek, 2017) (Zihni, et al., 2020) (Roscher, Bohn, Duarte, & Garcke, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Gunning & Aha, 2019) (Arun, 2020) (M, V, S, & H, 2020) (Ceni, Ashwin, & Livi, 2020) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Du, Liu, & Hu, 2019)

→ **Explainable surrogate models** (Preece, 2018) (Rio-Torto, Fernandes, & Teixeira, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Käde & Von Maltzan, 2019) (Arun, 2020)

→ **Features importance** (Bikmukhametov & Jäschke, 2020) (Singh, Sengupta, & Lakshminarayanan, 2020) (Rio-Torto, Fernandes, & Teixeira, 2020) (Böhle, Eitel, Weygandt, & Ritter, 2019) (Arras, Horn, Montavon, Müller, & Samek, 2017) (Zihni, et al., 2020) (Roscher, Bohn, Duarte, & Garcke, 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Käde & Von Maltzan, 2019) (M, V, S, & H, 2020)

→ **Model combination** (Preece, 2018) (Bikmukhametov & Jäschke, 2020) (Rio-Torto, Fernandes, & Teixeira, 2020) (Käde & Von Maltzan, 2019) (Gunning & Aha, 2019) (Sheh & Monteath, 2018)

→ **Prototype selection** (Preece, 2018) (Rio-Torto, Fernandes, & Teixeira, 2020)

→ **Rules based explanations** (Rosenfeld & Richardson, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)

→ **Salient masks** (Rosenfeld & Richardson, 2019) (Carvalho, Pereira, & Cardoso, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020) (Böhle, Eitel, Weygandt, & Ritter, 2019) (Arras, Horn, Montavon, Müller, & Samek, 2017) (Zihni, et al., 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (M, V, S, & H, 2020) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)

→ **Sensitivity analysis** (Rosenfeld & Richardson, 2019) (Carvalho, Pereira, & Cardoso, 2019) (Böhle, Eitel, Weygandt, & Ritter, 2019) (Arras, Horn, Montavon, Müller, & Samek, 2017) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Gunning & Aha, 2019)

(M, V, S, & H, 2020) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)

→ **Textual justification** (Preece, 2018) (Rosenfeld & Richardson, 2019) (Gunning & Aha, 2019) (Du, Liu, & Hu, 2019)

→ **Visual techniques** (Preece, 2018) (Rosenfeld & Richardson, 2019) (Carvalho, Pereira, & Cardoso, 2019) (Rio-Torto, Fernandes, & Teixeira, 2020) (Böhle, Eitel, Weygandt, & Ritter, 2019) (Karlsson, Rebane, Papapetrou, & Gionis, 2019) (Zihni, et al., 2020) (Spinner, Schlegel, Schafer, & El-Assady, 2020) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Käde & Von Maltzan, 2019) (Gunning & Aha, 2019) (Sheh & Monteath, 2018) (Arun, 2020) (M, V, S, & H, 2020) (Du, Liu, & Hu, 2019)

Clustering

→ **Decision Tree based explanations** (De Koninck, De Weerdt, & vanden Broucke, 2017)

→ **Explainable surrogate models** (De Koninck, De Weerdt, & vanden Broucke, 2017)

→ **Model combination** (De Koninck, De Weerdt, & vanden Broucke, 2017)

→ **Rules based explanations** (De Koninck, De Weerdt, & vanden Broucke, 2017)

→ **Visual techniques** (De Koninck, De Weerdt, & vanden Broucke, 2017)

Decision trees

→ **Transparent model types** (Preece, 2018) (Rosenfeld & Richardson, 2019) (Ariza, Arroyo, Caparrini, & Segovia, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Pintelas, Livieris, & Pintelas, A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability, 2020) (Käde & Von Maltzan, 2019) (Gunning & Aha, 2019) (Sheh & Monteath, 2018) (Arun, 2020) (Du, Liu, & Hu, 2019)

Dimensionality reduction

→ none

Ensemble learning

→ **Decision Tree based explanations** (Rosenfeld & Richardson, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Du, Liu, & Hu, 2019)

→ **Explainable surrogate models** (Ariza, Arroyo, Caparrini, & Segovia, 2020) (Käde & Von Maltzan, 2019)

→ **Features importance** (Rosenfeld & Richardson, 2019) (Ariza, Arroyo, Caparrini, & Segovia, 2020) (Bikmukhametov & Jäschke, 2020) (Zihni, et al., 2020) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019) (Du, Liu, & Hu, 2019)

→ **Model combination** (Bikmukhametov & Jäschke, 2020) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)

→ **Rules based explanations** (Rosenfeld & Richardson, 2019) (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)

→ **Sensitivity analysis** (Rosenfeld & Richardson, 2019)

→ **Visual techniques** (Käde & Von Maltzan, 2019)

Graph theory algorithms

→ none

Linear models

→ **Transparent model types** (Rosenfeld & Richardson, 2019) (Ariza, Arroyo, Caparrini, & Segovia, 2020) (Carvalho, Pereira, & Cardoso, 2019) (Zihni, et al., 2020) (Pintelas, Liaskos, Livieris, Kotsiantis, & Pintelas, 2020) (Pintelas, Livieris, & Pintelas, A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability, 2020) (Käde & Von Maltzan, 2019) (Arun, 2020) (Du, Liu, & Hu, 2019)

Natural Language Processing

→ **Deep explanations** (Arras, Horn, Montavon, Müller, & Samek, 2017) (Gunning & Aha, 2019)

→ **Features importance** (Arras, Horn, Montavon, Müller, & Samek, 2017)

→ **Salient masks** (Arras, Horn, Montavon, Müller, & Samek, 2017)

→ **Visual techniques** (Arras, Horn, Montavon, Müller, & Samek, 2017) (Gunning & Aha, 2019)

Nearest neighbour models

→ **Transparent model types** (Rosenfeld & Richardson, 2019)

Probabilistic and statistical models

→ **Transparent model types** (Rosenfeld & Richardson, 2019) (Carvalho, Pereira, & Cardoso, 2019) (Pintelas, Livieris, & Pintelas, A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability, 2020) (Gunning & Aha, 2019) (Arun, 2020)

Social recommender systems

→ **Recommender systems explanations** (Alshammari, Nasraoui, & Sanders, 2019) (Hong, Akerkar, & Jung, 2019) (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019) (Bharadhwaj & Joshi, 2018) (Hou, Yang, Wu, & Yu, 2019) (Chen, Yan, & Wang, 2019) (Cheng, Chang, Zhu, Kanjirathinkal, & Kankanhalli, 2019)

→ **Textual justification** (Hong, Akerkar, & Jung, 2019) (Amal, Tsai, Brusilovsky, Kuflik, & Minkov, 2019) (Chen, Yan, & Wang, 2019)

→ **Visual techniques** (Hong, Akerkar, & Jung, 2019) (Chen, Yan, & Wang, 2019)

Social semantic web

→ none

Support vector machines

→ **Features importance** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016) (Arras, Horn, Montavon, Müller, & Samek, 2017)

→ **Model combination** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)

→ **Prototype selection** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)

→ **Rules based explanations** (Guidotti, et al., A Survey of Methods for Explaining Black Box Models, 2019)

→ **Visual techniques** (Belle, Calster, Huffel, Suykens, & Lisboa, 2016)

Appendix O. Questionnaire of expert group 1 experts' opinion survey

1. What components for consumer credit scoring model would you consider? *

please answer in any of the following formats:
- single components (e.g. age of credit applicant)
- categories of components (e.g. demographic features)
- framework for credit scoring (e.g. 5 C's of credit scoring)

Enter your answer

2. What components for consumer credit scoring model would you consider in case of receiving access to applicant's social media data? *

please answer in any of these formats:
- single components (e.g. user's home town)
- categories of components (e.g. user's posts)

Enter your answer

3. (optional) How would you rate the completeness of the following categorization (link below)?

https://1drv.ms/t/s!Am1ziNaywK90hfRxdftfMGY6f1-CyQ

☆ ☆ ☆ ☆ ☆

4. (optional) Please comment your answer to the question 3 (categorization completeness).

Enter your answer

5. What approaches to derive consumer credit scoring model components from social media data would you consider? *

please answer in any of the following formats:
- single approaches (e.g. SVM with linear kernel)
- categories of approaches (e.g. deep neural networks)

Enter your answer

6. (optional) How would you rate the completeness of the following categorization (link below)?

https://1drv.ms/t/s!Am1ziNaywK90hfRyPiEj9BuTGhQjQg

☆ ☆ ☆ ☆ ☆

7. (optional) Please comment your answer to the question 6 (categorization completeness).

Enter your answer

8. (optional) How would you rate the following relational structure (link below)?

https://1drv.ms/u/s!Am1ziNaywK90hfR0CnJXHHPi-RDbDg

☆ ☆ ☆ ☆ ☆

9. (optional) Please comment your answer to the question 8 (relational structure).

Enter your answer

Appendix P. Questionnaire of expert group 2 experts' opinion survey

1. What approaches to derive consumer credit scoring model components from social media data would you consider? *

please answer in any of the following formats:
- single approaches (e.g. SVM with linear kernel)
- categories of approaches (e.g. deep neural networks)

Enter your answer

2. (optional) How would you rate the completeness of the following categorization (link below)?

https://1drv.ms/t/s!Am1ziNaywK90hfRyPiEj9BuTGhQjQg

☆ ☆ ☆ ☆ ☆

3. (optional) Please comment your answer to the question 2 (categorization completeness).

Enter your answer

4. What explainability techniques for social media user profiling approaches would you consider? *

please answer in any of the following formats:
- single techniques (e.g. SHapley Additive exPlanations)
- categories of techniques (e.g. explainable surrogate models)

Enter your answer

5. (optional) How would you rate the completeness of the following categorization (link below)?

https://1drv.ms/t/s!Am1ziNaywK90hfRz2AGlox7R0QLzAg

☆ ☆ ☆ ☆ ☆

6. (optional) Please comment your answer to the question 5 (categorization completeness).

Enter your answer

7. (optional) How would you rate the following relational structure (link below)?

https://1drv.ms/u/s!Am1ziNaywK90hfR0CnJXHHPi-RDbDg

☆ ☆ ☆ ☆ ☆

8. (optional) Please comment your answer to the question 7 (relational structure).

Enter your answer