TU WIEN Informatics

Dynamic Sentiment Analysis for Measuring Media Bias

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Thomas Elmar Kolb, BSc
Matrikelnummer 01426167

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Ass. Mag.rer.nat. Dr.techn. Julia Neidhardt

Wien, 22. März 2022

Thomas Elmar Kolb	Julia Neidhardt

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

TU WIEN Informatics

Dynamic Sentiment Analysis for Measuring Media Bias

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Thomas Elmar Kolb, BSc
Registration Number 01426167

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass. Mag.rer.nat. Dr.techn. Julia Neidhardt

Vienna, 22nd March, 2022

Thomas Elmar Kolb	Julia Neidhardt

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

Erklärung zur Verfassung der Arbeit

Thomas Elmar Kolb, BSc

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.

Wien, 22. März 2022

Thomas Elmar Kolb

Danksagung

Ich möchte mich sehr beim damaligen Dekan Prof. Hannes Werthner, für die Möglichkeit meine Masterarbeit in diesem Bereich zu schreiben, bedanken. Sein Engagement im digitalen Humanismus zeigt, wie wichtig dieser Forschungsbereich, in der sich heute schnell verändernden Welt, ist.

Besonders bedanken möchte ich mich außerdem bei meiner Betreuerin Dr. Julia Neidhardt für die umfassende Betreuung und Hilfe beim Verfassen meiner Arbeit.

Mein Dank gilt auch allen Projektpartner:innen an Universität Wien sowie der Akademie der Wissenschaften (ACDH-CH), mit welchen wir das DYSEN Projekt erfolgreich durchgeführt haben. Ohne die guten Zusammenarbeit, wäre die Durchführung dieser Arbeit, eingebettet in das DYSEN Projekt, nicht möglich gewesen.

Ebenfalls danke ich meiner Familie, für die Unterstützung, in der aktuell durch Corona geprägten Zeit. Ich möchte mich bei meiner Mutter und meinem Vater für die Möglichkeit an einer Universität zu studieren herzlich bedanken.

Acknowledgements

Special thanks goes to the former dean, Prof. Hannes Werthner, for the opportunity to write my Master's thesis in this field. His commitment to digital humanism shows how important this field of research is in today's rapidly changing world.

I am thankful for my supervisor Dr. Julia Neidhardt for her comprehensive supervision and help in writing my thesis.

Furthermore, I would like to thank all project partners at the University of Vienna and the Academy of Sciences (ACDH-CH), with whom we successfully carried out the DYSEN project. Without the good cooperation, the realisation of this work, embedded in the DYSEN project, would not have been possible.

Thanks to my supportive family for their support during the current corona period. Thank you to my mother and father for giving me the opportunity to study at an university.

Kurzfassung

Die Sentiment Analyse von Texten im Bereich der sozialen Medien und Nachrichten ist aktuell Gegenstand umfangreicher Forschung. Es ist ein weitbekanntes Problem, Maschinen zu lehren, wie die Stimmung von Texten, wie z.B. Nachrichten, automatisiert ausgewertet werden kann. Diese Masterarbeit hat das Ziel, dieses Konzept, im Bereich von Personen des öffentlichen Interesses, zu untersuchen. Im Speziellen werden Politiker_innen mit dem Fokus auf Wien betrachtet. Ein besonderes Interesse liegt außerdem auf dem Verlauf über die Zeit und die verschiedenen Medien. Sentiment Analyse wurde im Bereich der sozialen Medien und Nachrichten bereist umfangreich erforscht, nichtsdestotrotz sind aktuell noch viele Probleme ungelöst. Dies ist insbesondere der Fall für den Bereich der Nachrichten, welche wesentlich schlechter erforscht sind, als soziale Medien wie z.B. Twitter. Die Durchführung von Stimmunganalyse ist weiters stark sprachabhängig, wobei hier auch insbesondere das österreichische Deutsch einen Unterschied zu der standardmäßig durchgeführten Forschung im Standardhochdeutsch darstellt. Um diese Forschungslücken zu untersuchen, wird in dieser Arbeit eine Analyse mittels verschiedener Methoden des überwachten maschinellen Lernens durchgeführt. Diese beleuchtet den Bereich der Sentiment Analyse im Bezug zu Politiker_innen über die Zeit und den verschiedenen Nachrichtenquellen. Als Datenquelle wird der Austrian Media Corpus (AMC) herangezogen. Es hat sich gezeigt, dass Sentiment Analyse in dieser speziellen Domäne sehr herausfordernd ist. Nichtsdestotrotz zeigen die Ergebnisse, dass es möglich ist die Polarität von Politiker_innen über die Zeit zu bestimmen. Moderne Algorithmen, welche dem letzten Stand der Technik entsprechen, wie z.B. BERT basierte Modelle, übertreffen traditionelle Methoden. Auf BERT basierende Modelle haben aber den Nachteil, dass sie nicht transparent sind. Um diesen Nachteil auszugleichen, wurde zusätzlich ein lexikalischer Ansatz erfolgt, wodurch ein neues Stimmungswörterbuch entstanden ist. Dieses Sentiment Wörterbuch ist öffentlich zugänglich und kann für weitere Forschung benutzt werden. Das Sentiment Wörterbuch hat den Namen “Austrian Language Polarity in Newspapers (ALPIN)”. Die entwickelten Modelle bilden die Grundlage für eine Webanwendung zur Erforschung der Medienberichterstattung über Wiener Politiker_innen und der damit verbundenen Stimmungsdynamik.

Abstract

Analyzing the sentiments of texts in the field of social news and news media is a big area of interest for many researchers around the world. It is a well-known problem to "teach" machines to understand the sentiments of texts e.g. news media. This master thesis aims to unveil the sentiments towards persons of public interests, who are often presented in emotionally charged contexts, for different media and over time with a focus on Vienna. Although sentiment analysis has been widely applied for analysing news and social media content, there are still many challenges unsolved. This is particularly true for the area of news media as it is not as well researched in this context as the area of social media (e.g., Twitter). Sentiment analysis is strongly language dependent. Sentiment analysis of German texts, however, hardly considers the specifics of Austrian German. To tackle these research gaps, this research performs a supervised machine learning (SML) analysis for analyzing the sentiment towards politicians over the time, and different media of the Austrian Media Corpus (AMC) different methods were compared. Sentiment analysis has been shown to be very challenging in this narrow domain. Nevertheless, results show that it is possible to predict the polarity of politicians over time. Modern state-of-the-art approaches such as BERT based models outperform traditional approaches but are not transparent, which is important when it comes to explainability and fairness. To overcome this lack of transparency, a lexical-based method was used, resulting in a new sentiment dictionary. This sentiment dictionary can be further used for research in this field and is called "Austrian Language Polarity in Newspapers (ALPIN)". The developed models form the basis of a web application to explore media coverage of Viennese politicians and related sentiment dynamics.

Contents

Kurzfassung **xi**

Abstract **xiii**

Contents **xv**

1 Introduction **1**
1.1 Problem Statement and Motivation 1
1.2 Research Questions 2
1.3 Expected Results 2
1.4 DYSEN Project and Web Application 2
1.5 Methodology 3
1.6 Structure of the Work 5

2 State-of-the-Art **7**
2.1 Political Polarization: What Is "Polarization"? 7
2.2 Media Bias: What Is the Definition of "Media Bias"? 8
2.3 Sentiment Analysis 9
2.4 Austrian German Language in News Media: Why Is This "Special"? . 12

3 Data **13**
3.1 Austrian Politicians: What Are "Austrian" Politicians? 13
3.2 Politicians Archive (POLAR) 13
3.3 Austrian Media Corpus (AMC) 14
3.4 One Million Posts Corpus 14
3.5 Austriacisms 14

4 Preprocessing **15**
4.1 Tooling 15
4.2 Selection of Viennese Politicians 16
4.3 Austria Media Corpus (AMC) 17
4.4 STANDARD Data-set 25
4.5 Austriacisms 27

5 Data Annotation **29**
5.1 Inter-Rater Reliability . . . 29
5.2 AMC . . . 30
5.3 Austriacisms . . . 32

6 Lexicon Based Approach **39**
6.1 Methodology . . . 40
6.2 AMC Data-set . . . 41
6.3 STANDARD Data-set . . . 41
6.4 AMC & STANDARD Data-set . . . 41
6.5 Austriacisms . . . 42
6.6 Postprocessing . . . 46
6.7 Evaluation . . . 48

7 Supervised Approaches **55**
7.1 Tooling . . . 55
7.2 Methodology . . . 55
7.3 Results & Evaluation . . . 59

8 Discussion **63**
8.1 Preprocessing & Data Annotation . . . 63
8.2 Lexicon Based Approach Vs. Supervised Approaches . . . 64
8.3 Distribution of Politicians . . . 64
8.4 Web Application . . . 66
8.5 Ethical Questions . . . 69

9 Conclusion **71**
9.1 Summary . . . 71
9.2 Contribution . . . 73
9.3 Future Work . . . 74

List of Figures **75**

List of Tables **77**

Bibliography **79**

CHAPTER 1

Introduction

1.1 Problem Statement and Motivation

Austria has a broad and diverse news media landscape[1], which includes magazines, television and many newspapers. The news media are covering different regions and target groups and are also owned by different organizations, companies and persons.

Due to the work of politicians they are often shown in an emotionally charged "polarized" context. This leads to potentially varying positive or negative representations in different media as well as over time. It can be assumed that if a politician is involved in a scandal, there is an evidence of media "bias" when performing sentiment analysis of news articles. There is also the possibility that there is a consistent attitude towards politicians of the same party. This thesis aims to make such biases transparent and visible. To address these hypotheses, it is required to enhance the state-of-the-art in sentiment analysis applied to the domain of politician in news media. The German language especially the Austrian German is not as much used in the area of sentiment analysis as the English language [Sid19, BS16]. In addition, many models are built upon English language models therefore, they are not directly applicable to the German language. Sentiment Analysis is generally very domain dependent. In this work there is also the additional dynamic setting of time and different media which adds an additional layer of complexity to the models.

[1] https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/

1.2 Research Questions

Based on the aim of this work the research questions are defined as:

RQ1: To what extent is it possible to predict the polarization of politicians over time in different media?

RQ2: How well do different approaches of machine learning perform in predicting the polarisation of politicians in the context of sentiment analysis in the Austrian news media?

1.3 Expected Results

The main objective of this master thesis is to develop machine learning and lexicon based methods to measure the emotional polarization of persons of public interest in news media over time and for different media. To limit the scope only persons of public interest in the area of Vienna are considered. As a data-set the austrian media corpus (AMC)[2] created by the work of Ransmayr et al. [JKM17] is used. This corpus covers almost the entire Austrian media landscape i.e. all newspapers and weekly magazines of more then the last 30 years and it contains over 40 million articles.

1.4 DYSEN Project and Web Application

This thesis is part of the project "Dynamic Sentiment Analyses As Emotional Compass for the Digital Media Landscape" (DYSEN)[3] which is funded by the city of Vienna (grant number: MA7-737909/19).

The focus of this work, which is embedded into the DYSEN project, is the extraction, labelling, transformation, prediction (by using machine learning (ML) based methods as well as lexicon based approaches) and evaluation of the data. The linguistic part and the crowd sourcing planning and execution on the platform SoSci[4] and Prolific[5] was done by the project colleagues of the University of Vienna. The resulting website, which displays the by this work calculated sentiment scores, were programmed by the colleagues of the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)[6]. The DYSEN application[7] shown in figure 1.1 shows the web application which was created during the project and is freely available for usage.

[2] https://amc.acdh.oeaw.ac.at/
[3] https://dylen.acdh.oeaw.ac.at/dysen/
[4] https://www.soscisurvey.de/
[5] https://www.prolific.co/
[6] https://www.oeaw.ac.at/acdh/acdh-ch-home
[7] https://dysen-tool.acdh.oeaw.ac.at/

Figure 1.1: DYSEN web application

1.5 Methodology

Design science research is as proposed by Hevner [Hev07] a framework for assessing research in computer science and information systems. It also is a model which guides the researcher in doing the research in a structured way. Hevner [Hev07] proposed that Design science research can be described with three related circles of activities:

- Design cycle: build design artifacts & processes, evaluate
- Rigor cycle: grounding, additions to knowledge base (KB)
- Relevance cycle: requirements, field testing

1.5.1 Design Cycle

The design cycle retrieves on the one hand the requirements from the relevance cycle and on the other hand the state-of-the-art from the rigor cycle. In this master thesis different algorithms need to be applied on the given data-set to check which of the methods are performing best on the given domain. Evaluation of the artifact is done ex post in an artificial way.

The following steps are required to be performed in the design cycle:

- Data selection (see section 3)
- Preprocessing & data annotation (see section 4 & 5)
- Identifying the best fitting method & approach (see section 6 & 7)
- Evaluation and comparison (see section 8 & 9)

1.5.2 Rigor Cycle

The rigor cycle ensures that the thesis is innovative, relying on the state-of-the-art and based on past knowledge. The contributions of the research need to flow back to the knowledge base.

Literature Review

To define the state-of-the-art in the area of sentiment analysis a literature review was performed. The current state-of-the-art is described in chapter 2.

Contribution

The publication of the research results is done on the one hand through this master's thesis. On the other hand, a publication has been prepared and submitted to LREC 2022 conference, where it is currently under review. In addition, the language resource created is available online. More details can be found in the last chapter in the "Contribution" section 9.2.

1.5.3 Relevance Cycle

As stated by Hevner [Hev07] the desire of the relevance cycle is to improve the environment. This thesis aims to improve sentiment analysis applied to Austrian German in the domain of the Austrian news media. Therefore several issues need to be tackled e.g.:

- Austrian German language which is not so well researched as the English language.
- Analysis of news media which is also not so well researched as the analysis of e.g. tweets (Twitter).
- Dynamic setting of analysing the change of polarization of politicians over time.

1.6 Structure of the Work

This work is structured as follows: after chapter 1 the state-of-the-art is briefly explained in chapter 2. The 3rd chapter explains the data sources which were used. To understand how they were preprocessed the corresponding steps are explained in the 4th chapter. In the 5th chapter the data labeling is outlined. The 6th chapter is about the lexicon based approaches. To compare them the 7th chapter shows different supervised approaches which can be applied in this domain. The results are discussed in chapter 8 which leads into a conclusion in the last the 9th chapter.

Figure 1.2 displays the methodological approach in a compact form. The green pillar is referring to the machine learning (ML) based approach, the blue and red to the lexicon based approach. This leads to different ML models and a newly created sentiment dictionary called "Austrian Language Polarity in Newspapers (ALPIN)".

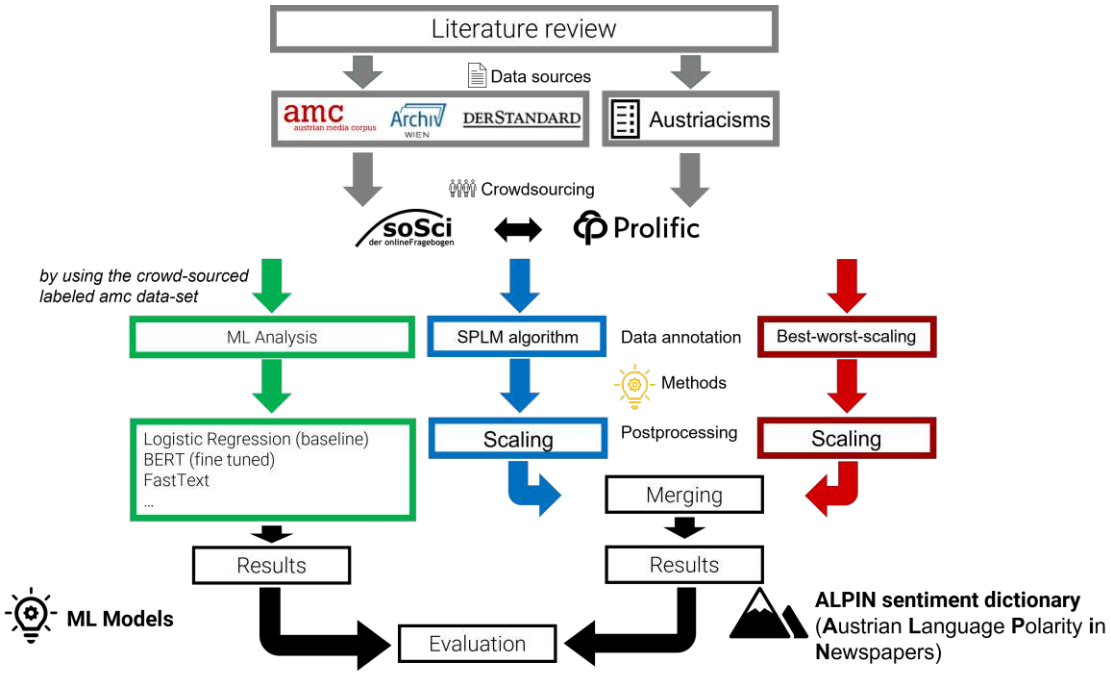

Figure 1.2: workflow

CHAPTER 2

State-of-the-Art

In this chapter the state-of-the-art and the related work is briefly presented. Relevant aspects in this research area are:

- Political polarization: what is polarization?
- Media bias: what is the definition of media bias?
- Sentiment analysis: what is the definition of sentiment analysis?
- The methods: what models can be used for sentiment analysis?
- The domain: what is special about Austrian German in the area of politics and news papers?

2.1 Political Polarization: What Is "Polarization"?

In Fiorina & Abrams [FA08], mass polarization and elite polarization are mentioned as important for the study of polarization in the political domain. They state that political polarization is a wide-ranging and polarized topic. Elite polarization is described by polarization of the governmental party against the opposition party [BC18].

Mass polarization is a meta term which covers a set of different polarization classifications e.g. Ideological Consistency, Ideological Divergence, Perceived Ideological Polarization and Affective Polarization [Lel16]. Lelkes [Lel16] describes the four different polarization classifications as:

- Ideological consistency: *"the degree to which people consistently align themselves with one side or another"* [Lel16] this statement is based on the work of [DDKO14, FA08].

- Ideological divergence: *"To Fiorina and colleagues, the mass public is polarized if the distribution of responses on this scale is bimodal"* [Lel16]. The mentioned scale is called 7-point liberal-conservative scale and is described as consistency scale to identify ideological alignment.
- Perceived ideological polarization: *"perceived ideological polarization, or the degree to which the mass public perceives the parties and their followers to be polarized"* [Lel16] which is based on multiple works of different researchers in the political area.
- Affective polarization: *"'feeling thermometers', measures that ask respondents to indicate on a 101-point scale how warm or cold they feel toward each party, and substract feelings toward the out-party*[1] *from feeling toward the in-party*[2] *"* [Lel16].

In Haselmayer et al. [HWM17] it is shown that a partisan bias in news media in the Austrian area exists. In this thesis, "political polarization" and the term "polarized context" are related to the question:

Is a politician portrayed by an editor in a particularly positive or negative way?

It is assumed that an editor can decide for himself or herself whether he or she wants to write about a particular politician. This decision can be made consciously or unconsciously. Its common knowledge that media are "ideological institutions" and therefore can never be 100% objective, as described in Herman & Chomsky's [HC10] propaganda model. Another assumption is that if the politician belongs e.g. to the left wing and the print media is also on the liberal side, they could decide not to write about left wing politicians in a negative way. All these different reasons are influenced by the polarization described in section 2.1.

2.2 Media Bias: What Is the Definition of "Media Bias"?

This section outlines the term "media bias" and the different sub-types as described by Eberl et al. [EBW17]. To understand media bias, one must know how bias is defined. Bias is the opposite of a neutral and objective news article or statement about a politician. Bias in news-media is another umbrella term that can be broken down into a more specific subset: Visibility, Tonality and Agenda. Visibility describes how often a particular politician appears, tonality describes how positively or negatively the media reports and agenda is defined as the media's decision to report or not report about a topic of a particular politician.

This work highlights all three different types of media bias. For visibility, the relative frequency of a politician is statistically calculated. Tonality is represented by the

[1] Refers to a different not the political party with which one identifies [Lel16].
[2] Refers to the political party with which one identifies [Lel16].

calculated sentiment score for each politician, and agenda is represented by comparing different politicians to show an imbalance between different political directions.

2.3 Sentiment Analysis

Sentiment analysis is a rapidly developing field of research. Especially since the increase in computational possibilities, this field has developed strongly over the past two decades [MGK16, BKKK16]. In one of the earlier works in this research area by Nasukawa et al. [NY03] sentiment analysis is described as the ability to positively or negatively evaluate the polarity of sentiment and the corresponding opinion on a given subject. Please see below the description what the term "Dynamic sentiment analysis" (subsection 2.3.1) in this work stands for, the various methods and algorithms used to perform sentiment analysis (subsection 2.3.2) and how they relate to the field of news media (subsection 2.3.3).

2.3.1 Dynamic Sentiment Analysis

One might ask why dynamic sentiment analysis? This thesis carries the term "dynamic" in its name because the sentiment scores of politicians are calculated across time and media. The settings are dynamic, as the sentiment scores can also adjust dynamically across parameter combinations.

2.3.2 Algorithms and Methods Used in Sentiment Analysis

There is a wide range of different algorithms that researchers apply to determine a sentiment from given input data [MHK14, vAvdVB21, Liu15, DMGDlP20, YV20]. In this research, the supervised and dictionary based approaches are particularly important. Nevertheless, there are other approaches such as unsupervised approaches and others that are used for sentiment analysis. In the upcoming paragraphs the two different types of machine learning, supervised machine learning (SML) and lexicon based approaches (LBA), are briefly explained.

Supervised Approaches

SML is a meta term which also includes a very recent area of research called deep learning (DL). Van Atteveldt et al. [vAvdVB21] describes the history of SML and DL very detailed. In the early days sentiment analysis research started by utilizing rule based methods which are there described as "codebook". For example in the referenced paper of Aday [Ada10] manual annotation and coding were performed. The importance of rule based approaches is also shown in Langleys paper *"The Changing Science of Machine Learning"* [Lan11]. Later more advanced algorithms were proposed e.g. Naive Bayes (NB) and Support Vectore Machines (SVM) [vAvdVB21]. In the last years there is a trend into the direction of DL which currently yields to better results as the traditional SML methods. In Dang et al. [DMGDlP20] and Yadav & Vishwakarma [YV20] recent DL methods are

described. They highlight Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and many more. The goal for DL based methods is to learn the domain-language. This allows to additionally answer out-of-scope examples during performing classification tasks. In addition to the already mentioned DL methods "Bidirectional Encoder Representations from Transformers (BERT)" proposed by Devlin et al. [DCLT19] counts to the most advanced and state of the art algorithms in the area of sentiment analysis. This is shown by the publication of Hoang et al. [HBR19] where they use aspect-based sentiment analysis using BERT.

Lexicon Based Approaches (LBA)

Lexicon based approaches are another important research area in performing sentiment analysis. They are often used in the linguistics area. Although they are outperformed by deep learning systems they continue to be used in research [vAvdVB21]. One important disadvantage of DL systems ist that they are not transparent and explainable. One could also state that most of the DL systems are a "black box" [XGR20]. LBA in comparison is explainable by default and allows further analysis why a certain algorithm is classifying in a specific way [vAvdVB21].

The various subcategories of LBA can be defined in different ways. A definition by Medhat et al. [MHK14] divides the domain of lexicon-based methods into dictionary-based methods, corpus-based methods, lexicon-based and natural language techniques. They describe that dictionary-based methods are created by manually labeling a representative part and extending it with synonyms and antonyms based on larger text corpora. One disadvantage they describe is that it is not possible to create a domain-specific language resource. Domain dependency is lost by enriching with data from a larger text corpora. On the other hand, in Medhat et al. [MHK14] corpus-based methods are described as capable of representing opinion words with their context. A text corpus does not just show a word with its associated sentiment score, as a sentiment dictionary would. Last, but not least, the subcategory *"lexicon-based and natural language techniques"* [MHK14] is mentioned. This category describes that often additional natural language techniques (NLP) are needed to extract the context of an opinion word by using statistical methods or machine learning based models to obtain e.g. named entities (NE), part of speech tagging (PoS) and more.

In a very recent publication of Catelli et al. [CPE22] a comparison of lexicon-based and BERT based methods in the field of sentiment analysis is shown. They show that BERT based methods achieve a higher f1 score and accuracy value than lexicon-based approaches. In their analysis, they describe that lexicon-based methods are particularly good when only a small data-set is available. In their conclusion, they additionally mention that the disadvantage of BERT based methods is the need for computational resources and that there is potential to further improve the lexicon-based approach used in their publication.

Weakly Supervised Approaches

It is often not possible to perform a sentiment analysis task with strong supervision. The concept of weakly supervised approaches is described in detail by Zhou [Zho17]. Weakly supervised learning is classified into three different categories: incomplete, imprecise and inaccurate supervision. Incomplete supervision is described as a small number of labels with a large number of unlabeled data. The paper highlights two methods to overcome this problem: first, having an expert who can be used to increase the amount of ground-truth labels *"active learning"* [Zho17]; second, performing a special version of semi-supervised learning, which Zhou [Zho17] calls *"transductive learning"*. Inaccurate supervision, mentioned in the publication, is described as the presence of features to predict a particular task, but not the complete information to train the learner optimally. The last variant, by Zhou [Zho17] called, *"imprecise monitoring"*, is described as the presence of labels that are known not to correspond to the exact ground truth. This is explained, for example, through noise in the data.

2.3.3 Sentiment Analysis in News Media

Sentiment analysis can be applied on a broad range of different texts and text forms. It is currently very widely applied to analyse short texts e.g. Twitter[3] data and ratings e.g. IMBD[4] movie reviews [KJ20, BKBH21, Sid19]. This references describe that there is a wide range of methods applied on this type of data especially for the English language. On the other side non English languages and sentiment analysis performed in the domain of news media are not so well researched. This is shown by a publication [KBK+21] of the DYSEN project (see section 1.4) into which this thesis is embedded. Pereira [Per21] is also describing a research gap, if English is compared with non English languages, in this case, the Portuguese language.

The publication of Haselmayer & Jenny [HJ17] and Rudkowsky et al. [RHW+18] is part of a paper series where they applied sentiment analysis in the area of parliamentary speeches in the Austrian German area. Since this thesis is focusing on the area of news media in the political domain. The topic is related but not exactly matching the domain of this work. Backfried & Shalunts [BS16] analysed the refugee crysis in news media of the German, Austrian and Swiss area. They used an lexical based approach (SentiSAIL), whereby they also stated, that the available methods in sentiment analysis for the German language are limited. Souma et al. [SVA19] researched news sentiment analysis by using deep learning methods. They aimed to predict the market performance by forcasting financial news sentiments. Methods used by this approach are GloVe which is a global vectors for word representation in combination with a polarity classification into positive or negative by stock price changes. Deep learning Recurrent Neural Networks (RNN) composed of Long Short-Term Memory (LSTM) units are used to perform the deep learning process. They state in their conclusion that there is potential to further improve

[3] https://twitter.com
[4] https://www.imdb.com/

the results by varying the methodology by using different methods e.g. Convolutional Neural Networks (CNN), seq2se, or attention based models. This results are very valuable despite the data used for the analysis is mainly English.

2.4 Austrian German Language in News Media: Why Is This "Special"?

The focus of this thesis is, as already suggested by the title, on Austrian German. Depending on the used methods it is important to keep in mind that German is spoken in different variations over Germany, Austria and the German speaking part of Switzerland [Amm95]. Lexical difference as described by Ammon [Amm95] is important if lexical-based methods e.g. dictionary-based approaches are utilized. They heavenly depend on the sentiment assigned to words. Including regional words can improve the quality of a created language resource by increasing the diversity of the resulting language resource.

The second difference to the commonly performed sentiment analysis is the domain of news media. It's articles are quite different from a movie review or a short tweet. They are often written in a neutral way and most of the times way longer than a single tweet. This is also described by the work of Raina [Rai13].

CHAPTER 3

Data

This chapter describes the data sources used to answer the research questions. This chapter provides a detailed insight into the selection criteria for the "Austrian" politicians (section 3.1). Various data sources are used with the Politician Archive (POLAR) (section 3.2) and the Austrian Media Corpus (AMC) (section 3.3) being relevant to all approaches. The One Million Posts Corpus (section 3.4) and the Austriacisms (section 3.5) are used to perform the lexicon-based approach presented in chapter 6.

3.1 Austrian Politicians: What Are "Austrian" Politicians?

As described in section 1.4 this master-thesis is embedded into a research project funded by the city of Vienna. The scope is limited by the DYSEN project to politicians related to the city of Vienna. With the methodology described in this thesis generalization is possible with less effort. The term "Austrian politicians" is defined as politicians related to the city of Vienna which is further described in the preprocessing subsection 4.2. The list of politicians is collected by using the POLAR database described in the upcoming section 3.2.

3.2 Politicians Archive (POLAR)

The archive of Viennese politicians (POLAR)[1] in German "Politikerinnen und Politiker Archiv" is a politician database provided by the city of Vienna which contains more than 1200 Viennese politicians. For each of the politicians biographical details and political functions are listed. The POLAR contains information about Viennese politicians starting with 1918 up to the current year.

[1] https://www.wien.gv.at/kultur/archiv/politik/

3.3 Austrian Media Corpus (AMC)

The Austrian media corpus (AMC)[2] is one of the largest German text corpora covering almost the entire print media landscape[3] of Austria. With approximately 45 million articles collected over more than 30 years, this corpus represents a comprehensive collection for research. The raw text data provided by the Austria Press Agency (APA)[4] were enriched with linguistic data on several levels by the research project of Ransmayer et al. [JKM17]. The enriched form of the news media texts provided by the AMC corpus by Ransmayer et al. [JKM17] is further used in this work. The current latest version of the corpora v3.1 is used for performing the data extraction in section 4.3. This corpus is available[5] for research and teaching for everyone after registering and accepting the terms of use.

3.4 One Million Posts Corpus

The corpus of Schabus et al. [SST17] also called the "One Million Posts Corpus" is a labeled data-set of user posts posted on the STANDARD news media website[6]. The items were annotated by experienced human annotators. The focus was on labeling negativity, resulting in very few positive posts in the dataset. This dataset is further referred by the term "STP" and available at GitHub[7]. The corpus contains in total 1.011.773 posts whereby 11.773 are labeled. The labeled categories are: Sentiment, Off-Topic, Inappropriate, Discriminating, Feedback, Personal Stories and Arguments Used which is described in detail in the referenced publication.

3.5 Austriacisms

The term "Austriacism" refers to a variant of a word used in a particular region or area. In this case, the country Austria. For a comparative overview, see Ammon et al. [ABE16] in the book "Variantenwörterbuch des Deutschen". In this work, the data from the referenced book are used in combination with an Austriacism list of Wikipedia[8]. The Wikipedia list is selected in addition to the book due to the fact that the Wikipedia list is more up-to-date and therefore can supplement the book's list with newer words.

[2] https://amc.acdh.oeaw.ac.at/
[3] https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/
[4] https://apa.at/
[5] https://amc.acdh.oeaw.ac.at/access-conditions/
[6] https://www.derstandard.at/
[7] https://ofai.github.io/million-post-corpus/
[8] https://de.wikipedia.org/wiki/Liste_von_Austriazismen

CHAPTER 4

Preprocessing

This chapter shows in detail which preprocessing steps were necessary for the analysis of the data. In chronological order, the selection of Viennese politicians is first described in section 4.2. Based on the created list of Viennese politicians, the preprocessing and extraction of the AMC data was performed, which is explained in section 4.3. For the lexicon-based approach, the preprocessing of the *One Million Posts Corpus* is outlined in section 4.4 and the creation of the Austriacism list is described in section 4.5.

4.1 Tooling

Preprocessing was performed by using the programming language Python with the corresponding packages to perform data-science tasks. The data retrieval steps are described in subsection 4.1.1 and the Part of Speach (PoS) tagging & tokenization in subsection 4.1.2.

4.1.1 Data Retrieval

Pandas [tea20] was used to perform data analysis and manipulation, BeautifulSoup[1] to parse the XML files given by the AMC and the package multiprocessing[2] to utilize more than one core. More specific frameworks are mentioned in the respective sections.

4.1.2 Part of Speach (PoS) Tagging, Tokenization and Lemmatization

For performing natural language processing (NLP) steps the "Natural Language Toolkit" (NLTK) [BKL09] is used. Part of Speach (PoS) tagging and lemmatization is performed by utilizing the spaCy framework developed by Honnibal et al. [HMVLB20]. The

[1] https://www.crummy.com/software/BeautifulSoup/
[2] https://docs.python.org/3/library/multiprocessing.html

tag to WordNet® word-form assignment is shown in the corresponding section of the preprocessing steps. The WordNet® lexicon itself is not used any further, only the word forms are "borrowed" from there to have a common style of word form assignment. WordNet® is a lexical database for the English language.

The overall analysis was performed on a server provided by the team of ACDH-CH. The hardware of the server was: 18 cores, 70GB ram, SSD storage and a NVIDIA A100 GPU. The AMC corpus (section 3.3) was directly mounted onto the server to allow access on the raw data of the corpus.

4.2 Selection of Viennese Politicians

To be able to perform a sentiment analysis in the Viennese area in the field of news media and politics, it was necessary to extract domain-specific texts from the AMC. In order to identify paragraphs and areas that refer to a specific Viennese politician, a Viennese politician list was created.

The list of Viennese politicians is created by using the the politician archive of Vienna POLAR (section 3.2) from the Vienna City and State Archives. The AMC (section 3.3) proposed by Ransmayr et al. [JKM17] contains media from the year 1986 to 2018. The result list is limited to politicians which were active in the timeframe between 1986 to 2020. Therefore, the politicians who were active between the 13th and the 20th parliamentary term were selected.

The list of politician's contains the following political functions:

- All members of the Vienna City Council and members of the Vienna State Parliament
- All members of the Vienna City Senate and the Vienna State Government

Two lists were extracted out of the dataset:

Politicians list with party: Contains duplicates if a politician was active in multiple parliamentary terms.

- Structure: ["Name","Link","Name before marriage","Party","Parliamentary term"]

Politicians (distinct): List of distinct politicians

- Structure: ["Name"]

The resulting unique list consists of 487 politicians related to the Vienna area which is shown in table 4.1.

	full name
0	Adolf;Aigner
1	Eveline;Andrlik
2	Josef;Arthold
3	Dolores;Bauer
4	Helmut;Braun
...	
482	Michael;Stumpf
483	Thomas;Weber
484	Christoph;Wiederkehr
485	Markus;Wölbitsch-Milan
486	Ernst;Woller

Table 4.1: Unique politician list based on POLAR after filtering.

docsrc	*type*	*from*	*to*	*region*	*article-count*	*token-count*
APA	agentur	1986	2018	agesamt	5.874.910	1.739.342.744
OTS	agentur	1989	2018	agesamt	1.695.514	663.153.108

Table 4.2: APA & OTS press releases based on AMC v3.1. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17].

4.3 Austria Media Corpus (AMC)

After creating a list of Viennese politician the extraction of paragraphs and text areas was performed. The structure of the AMC created by Ransmayr et al. [JKM17] is explained in detail in the next subsection 4.3.1.

4.3.1 Corpus Structure

The corpus described on an meta level in section 3.3 is the most important data-set, which is used in this work. Extraction and preprocessing of the corpus data was one very important step in this thesis. In the upcoming paragraphs the full media list is shown to outline how comprehensive the AMC is. Table 4.2 shows which press release media, Table 4.3 and 4.4 which news media and 4.5 which TV media transcripts are contained in the corpus. The different media sources “docsrc” are categorized according to their geographical region, which is indicated in the column “region”. The distinction is made according to the cardinal points “aost”, “amitte”, “asuedost” and “awest”. If the media source could not be assigned to a region, the classification “agesamt” is used. The categorization “specific” is used if the media source is not regional and/or refers to a specific topic.

docsrc	*type*	*from*	*to*	*region*	*article-count*	*token-count*
ACADEMIA	print	2008	2018	spezifisch	1.493	1.038.460
ARBEITW	print	2000	2018	spezifisch	6.127	4.974.496
AUGUSTIN	print	2003	2018	spezifisch	15.746	10.697.282
BAUERNZT	print	2010	2018	spezifisch	71.802	18.196.504
BVZ	print	2003	2018	aost	834.446	149.529.884
DATUM	print	2007	2018	agesamt	3.366	4.272.784
DIEWIR	print	2003	2018	spezifisch	6.095	3.595.779
ECHO	print	2004	2018	spezifisch	18.700	10.574.773
EMEDIA	print	2000	2018	spezifisch	36.446	11.780.249
FALTER	print	1998	2018	aost	110.566	56.313.323
FORMATDB	print	1998	2018	agesamt	131.940	45.099.435
FURCHE	print	1998	2018	agesamt	55.284	33.710.901
GEWINN	print	1998	2018	spezifisch	30.572	20.962.312
GRAZER	print	2008	2018	asuedost	43.870	7.184.280
HEUTE	print	2007	2018	aost	428.339	36.609.560
HOR	print	2000	2018	spezifisch	64.945	22.507.727
IM	print	1999	2018	spezifisch	15.360	7.710.408
KLEINE	print	1996	2018	asuedost	3.498.956	676.299.516
KONSUM	print	1996	2018	spezifisch	12.399	7.587.463
KRONE	print	1994	2018	agesamt	4.958.487	947.863.733
KTNMONAT	print	2005	2013	asuedost	8.364	3.134.528
KTZ	print	1999	2014	asuedost	646.192	122.010.288
KURIER	print	1992	2018	agesamt	3.019.003	854.255.840
KW	print	1996	2018	spezifisch	34.524	9.715.386
MEDIANET	print	2002	2018	agesamt	201.164	46.747.550
NEWS	print	1999	2018	agesamt	106.062	52.452.537
NOEN	print	1995	2018	aost	6.034.162	1.040.209.537
NVB	print	1997	2018	amitte	851.360	149.326.289
NVT	print	1997	2018	awest	648.903	166.128.689

Table 4.3: News media list based on AMC v3.1 part 1/2. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17].

docsrc	*type*	*from*	*to*	*region*	*article-count*	*token-count*
OBERRUND	print	2010	2018	awest	105.045	31.922.452
OEREICHE	print	2006	2015	agesamt	1.003.781	179.344.012
OOEN	print	1996	2018	amitte	1.633.293	411.252.621
PRESSE	print	1991	2018	agesamt	1.304.795	437.530.380
PROFIL	print	1994	2018	agesamt	110.977	74.478.738
SBGW	print	2007	2018	amitte	234.126	51.335.559
SN	print	1991	2018	agesamt	1.618.344	433.912.061
SOLI	print	2001	2018	spezifisch	3.993	1.177.686
SPORTZTG	print	1996	2018	spezifisch	49.761	19.175.202
STANDARD	print	1990	2018	agesamt	1.278.804	477.671.697
STMONAT	print	2005	2012	asuedost	7.991	2.788.583
SVZ	print	2007	2014	amitte	183.720	28.103.325
TREND	print	1994	2018	spezifisch	32.990	24.922.200
TT	print	1996	2018	awest	1.587.242	354.667.071
TTKOMP	print	2008	2018	awest	99.946	20.309.348
TVMEDIA	print	1999	2018	spezifisch	128.573	30.015.031
VN	print	1997	2018	awest	1.386.101	315.832.069
WIBLATT	print	1995	2016	agesamt	422.950	112.615.725
WIENER	print	2001	2018	spezifisch	14.474	7.155.223
WIENERIN	print	2001	2018	spezifisch	24.798	9.792.451
WOMAN	print	2003	2018	spezifisch	43.210	18.061.809
WZ	print	1996	2018	agesamt	794.771	304.950.055

Table 4.4: News media list based on AMC v3.1 part 2/2. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17].

docsrc	*type*	*from*	*to*	*region*	*article-count*	*token-count*
ATVVOLL	tv	2005	2018	agesamt	16.568	4.522.275
MWVOLL	tv	2003	2018	spezifisch	670.489	220.072.463
PRO7VOLL	tv	2007	2018	agesamt	4.322	1.243.372
PULSVOLL	tv	2007	2018	agesamt	9.831	2.819.171
SAT1VOLL	tv	2008	2018	agesamt	4.795	1.462.907

Table 4.5: TV media list based on AMC v3.1. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17]. Remark: “MWVOLL” is the Austrian public service broadcaster called “Österreichischer Rundfunk (ORF)”.

4.3.2 Excluded Media

As described in the beginning the thesis is based on Austrian news media around Viennese politicians therefore all news media which are not related to Vienna were excluded. In addition the television media and the APA & OTS press releases were excluded. APA & OTS press releases are mostly about press statements of certain organizations e.g. parties. The television media transcripts are also often about statements or interviews of politicians which does not fit into the target of the thesis to capture the media bias of news media.

The following media were excluded: APA, OTS, BAUERNZT, BVZ, ECHO, EMEDIA, GRAZER, HOR, IM, KLEINE, KONSUM, KTNMONAT, KTZ, KW, MEDIANET, NOEN, NVB, NVT, OBERRUND, OOEN, SBGW, SN, SOLI, SPORTZTG, STMONAT, SVZ, TT, TTKOMP, TVMEDIA, VN, WIBLATT, WOMAN, ATVVOLL, MWVOLL, PRO7VOLL, PULSVOLL, SAT1VOLL.

The resulting list after excluding the above mentioned news media contains: DIEWIR, STANDARD, WIENER, FALTER, WZ, AUGUSTIN, DATUM, FURCHE, KRONE, WIENERIN, NEWS, OEREICHE, PROFIL, KURIER, GEWINN, ACADEMIA, PRESSE, ARBEITW, FORMATDB, TREND, OESTERREICH, HEUTE.

4.3.3 Data Extraction

After excluding the non relevant media of the news media list the text items around Viennese politicians needed to be extracted. Therefore an area/phrase based extraction approach was chosen. Extraction is possible on multiple levels in the literature often document, paragraph and phrase are mentioned as possible text extraction levels. This is described in detail in the work of Balaji et al. [BNH17]. Document- and paragraph-based text extraction was not a viable solution because multiple politicians can occur in both a document and a paragraph. An assignment of the corresponding sentiment scores to a specific politician would've not been possible. The selected text extraction level in this work is phrase based text extraction. Aspect based sentiment analysis was not possible due to the higher financial impact on the labeling step which was no viable option for the project into which this thesis is embedded. Balaji et al [BNH17] also describes the concept of "Named Entity Extraction" which is very important for this thesis. In this work named entities are the first-, middle- and last-name of the politicians listed in the politician list. The AMC annotation structure is described in detail on the corresponding project website[3].

Phrase extraction in this thesis works as follows: If a politician is contained in a given paragraph, the paragraph is extracted. If this paragraph exceeds the maximum allowed length[4], the paragraph is shortened sentence by sentence, starting with the last sentence. If this is not sufficient, the sentences at the beginning of the paragraph are removed. In

[3] https://amc.acdh.oeaw.ac.at/dokumentation/korpusinhalt-attribute/

[4] The maximum item length is restricted to the copyright of the AMC

the end, if the remaining sentence containing the politician is longer than the allowed token count, this element is dropped and not used for further analysis. After each removed sentence, the token length of the remaining paragraph is checked. If the token length is less than the maximum token count, the algorithm terminates.

Relevant for the extraction of the phrases around Viennese politicians is the “ner” tag which allows to identify if the matching politician is tagged as “PERSON”, the “lempos” for checking if the detected word is a noun and the “iob” which allows to identify if a entity consists of multiple tokens e.g. first- and last-name. The maximum size of a phrase is limited due to copyright reasons to a value of around 60 tokens. If a certain extracted phrase is shorter as 15 tokens this phrase is excluded from further analysis. The algorithm was developed by removing the first sentence at the end of a paragraph, since news media often describe a particular action or event of a politician and then mention the name of that politician afterwards.

The extraction was performed by using the programming language and tools mentioned in the tooling subsection 4.1.1. The whole corpus was processed whereby the program was programmed in a way to run on all cores of the server in parallel (multiprocessing).

The structure of the AMC data is explained in detail on the AMC website[5]. The coarse structure is:

```
<file>
    <doc>
        <field>
            <p>
                <s>
                ...
                </s>
                ...
            </p>
            ...
        </field>
        ...
    </doc>
    ...
</file>
```

“<doc>” describes an article, “<field>” the title or headline of an article, “<p>” a paragraph, “<s>” an sentence. All mentioned tags except the “file” tag can occur multiple times in a single XML file of the raw data. One XML file is a specific news media e.g. FALTER on a specific day (one issue).

[5] https://amc.acdh.oeaw.ac.at/dokumentation/korpusinhalt-attribute/#Structs

Example

This is the first extracted item of the politician “Alexander Van der Bellen” of the news media “STANDARD” and the date “1990.11.28”:

Text: “*Der Faktor Arbeit werde ja vom Staat nicht absichtlich belastet, sondern nur weil es eben bequem sei, bei der Besteuerung an der Arbeit anzuknüpfen. Energieabgaben hätten aber eine Lenkungsabsicht - " Öko-Steuern erhöhen die Effizienz, denn sie korrigieren die Preise", argumentiert Alexander Van der Bellen.*”

Lemmata: “*die Faktor Arbeit werden ja von Staat nicht absichtlich belasten , sondern nur weil es eben bequem sein , bei die Besteuerung an die Arbeit anknüpfen . Energieabgabe haben aber eine Lenkungsabsicht - " Öko-Steuer erhöhen die Effizienz , denn sie korrigieren die Preis " , argumentieren Alexander Van die Bellen .*”

The original item contained more than the allowed 60 tokens. In this case due to the politician name is located on the end of the paragraph, the text first sentence was omitted.

Following data were extracted for further analysis for each occurrence of a matching politician:

- Politician name (full name of a politician)
- News media (e.g. Standard, Falter, ...)
- Date (e.g. 1990.11.28)
- Text reduced (based on maximum token count)
- Lemmata (based on text reduced)
- Text not reduced (full paragraph)
- Text with all taggings of the AMC itself
- Text reduced with PoS & negation tagging

Overall 494.111 items were extracted out of the AMC based on the list of Viennese politicians.

lemma, PoS and negation tagging
[[so, ADV, -], [eine, ART.Indef.Nom.Sg.Neut, -...
[[Roland, N.Name.Nom.Sg.Masc, -], [Sperk, N.Na...
[[feiern, VPP.Full.Psp, -], [werden, VFIN.Aux....
[[in, APPR.In, -], [die, ART.Def.Dat.Sg.Fem, -...
[[bei, APPR.Dat, -], [eine, ART.Indef.Dat.Sg.M...
...
[[bei, APPRART.Dat.Sg.Masc, -], [ÖVP-Parteivor...
[[Zukunftswerkstätte, N.Reg.Nom.Sg.Fem, -], [(...
[[die, ART.Def.Acc.Sg.Masc, -], [gepflanzt, AD...
[[in, APPRART.Dat.Sg.Masc, -], [Zug, N.Reg.Dat...
[[jetzt, ADV, -], [laufen, VFIN.Full.3.Sg.Pres...

Table 4.6: AMC politician phrases with lemma, PoS and negation tagging

tag	wn-type	short	description
ADJA	wn.ADJ	a	attributive adjectives
ADJD	wn.ADV	r	adjective with predicative or adverbial usage
ADV	wn.ADV	r	adverbs
N	wn.NOUN	n	noun
VFIN	wn.VERB	v	finite verb
VIMP	wn.VERB	v	imperative verbs
VINF	wn.VERB	v	infinitival verb
VPP	wn.VERB	v	participle verb

Table 4.7: Tag to WordNet® word-form assignment to harmonise the tagging

4.3.4 Part of Speech (PoS) Tagging & Lemmatization

The AMC already contains the PoS tagging[6] and lemmas[7] for each word in the entire corpus. Therefore, this data were extracted and prepared for the annotation trial described in chapter 5. The tagging is shown in table 4.6.

4.3.5 Mapping and Stop Word Removal

The extracted tagging were aligned to the over all data-sets common tagging schema which is shown in table 4.7. In the next step German stop words were removed with a list provided by the NLTK package. The preprocessed data-set after all the preprocessing steps is shown in table 4.8. Lemma without an assigned short-tag are not further used.

[6] https://amc.acdh.oeaw.ac.at/dokumentation/korpusinhalt-attribute/#posbase
[7] https://amc.acdh.oeaw.ac.at/dokumentation/korpusinhalt-attribute/#lemma

AMC phrases
[(Ansinnen, n), (scheinen, v), (Chef, n), (Bür...
[(Roland, n), (Sperk, n), („), (Vorsitzende,...
[(feiern, v), (Szenelokal, n), (",), (Stamper...
[(ÖVP, n), (sogenannt, a), (Westachse, n), (En...
[(Erfolg, n), (Volksbegehren, n), (Jänner, n),...
...
[(ÖVP-Parteivorstand, n), (oberösterreichisch,...
[(Zukunftswerkstätte, n), ((,), (I., n), („ ...
[(gepflanzt, a), (Baum, n), (Spitzenkandidat, ...
[(Zug, n), (EU-Programm, n), (",), (Urban, r)...
[(laufen, v), (parallel, r), (Studie, n), (MA,...

Table 4.8: AMC phrases based on Viennese politicians after preprocessing

4.3.6 Statistics

The corpus was further analysed to calculate the occurrence of the politicians over time and the different news media. This is especially important to answer RQ1 whereby this results are related to the concept of *visiblity* mentioned in the definition of *media bias* in section 2.2.

The following metrics were calculated:

- relativeFrequency group_by politician:
  - unit: per million
  - calculation: (hits of this query in all years and in all media sources / all tokens in all years and in all media sources) * 1.000.000
- relativeFrequency group_by politician and year:
  - unit: per million
  - calculation: (hits of this query in THIS year and in all media sources / all tokens in THIS year and in all media sources) * 1.000.000
- relativeFrequency group_by politician, year, media source:
  - calculation: (hits of this query in THIS year and in THIS media source / all tokens in THIS year and in THIS media source) * 1.000.000

A hit was counted if the following attributes were matching:

- nomen tagged with “-n”

column	description
ID_Post	Unique ID of a post
Headline	The Headline/Subject of a post
Body	Text of a post
Category	The sentiment labelled by the professional annotators (SentimentNeutral, SentimentPostive, SentimentNegative)
Text	Content of Headlinge + Body

Table 4.9: STANDARD posts data-set structure

- ner type = "Person"
- iob type = "B"
- Full name (first-, middle- and last-name) is used to match the politicians

The overall token count is defined as: all tokens except tokens inside areas which are marked with the XML attribute "dupl". In detail the tokens contained in each "doc" tag with the "field" tag and the attribute "title" or "inhalt" which are not tagged as "dupl" are counted.

4.4 STANDARD Data-set

The in this thesis used data-set is a reduced version derived from the corpus of Schabus et al. [SST17] which was created by the colleagues at University of Vienna. The new data-set contains only "ID_Post", "Headline", "Body", "Category" and "Text" which is described in table 4.9.

Table 4.10 gives a brief overview of the data contained in the data-set. For further analysis the columns "Text" and "Category" are used. The column named "Category" contains the sentiment assigned to the post. The sentiment score is converted to "positive", "neutral" and "negative" to have a common classification over all the different data-sets used in this thesis. In total there are 3599 rows whereby the sentiment classification is distributed into: 1865 neutral, 1691 negative and 43 positive entries.

4.4.1 Part of Speech (POS) Tagging & Lemmatization

For text tokenization, part of speech tagging and lemmatization the spaCy framework is utilized. The trained pipeline "de_core_news_sm"[8] were selected. Tokenization is different for the various languages therefore always a dedicated for the language trained

[8] https://spacy.io/models/de

	category	text
3326	SentimentNeutral	Top qualifizierte Leute verdienen auch viel.
5321	SentimentNegative	Gott sei dank ist für sie eine Umfrage alles, ...
5590	SentimentNeutral	" Die FPÖ wird aus allen Rohren schießen und d...
6015	SentimentNegative	Weil es dein meisten Leuten verständlicherweis...
8213	SentimentNeutral	Na wer weis was da vorgefallen ist...
...	...	...
1004115	SentimentNeutral	Russland ist in wk1 vorzeirig ausgestiegen. ;-...
1004189	SentimentNeutral	Was tendenziell kein schlechter Tausch wäre, w...
1004571	SentimentNeutral	Was? Unsinn! Der Linguistik turn beschränkt si...
1006462	SentimentNegative	wien verschreckt investoren, wenn sie trotz po...
1006960	SentimentNegative	Früher haben sie ein viertelteleFon beantragen...

Table 4.10: Overview of the STANDARD posts data-set

tag	wn-type	short	description
ADJ	wn.ADJ	a	adjective
ADV	wn.ADV	r	adverbs
NOUN	wn.NOUN	n	noun
PRON	wn.NOUN	n	proper noun
PROPN	wn.NOUN	n	proper noun
VERB	wn.VERB	v	verb

Table 4.11: Tag to WordNet® word-form assignment to harmonise the tagging

pipeline is provided by the framework. The “de_core_news_sm” pipeline is trained based on different German news media sources[9,10,11,12].

4.4.2 Mapping and Stop Word Removal

Table 4.11 shows the assignments of “tag” to “wn-type”. The preprocessed data-set is shown in Table 4.12 which displays the data-set after the various preprocessing steps. The tuples without an assigned word-form are not further used. Due to tagging only the in table 4.11 shown word types, punctuation and other not relevant words are automatically omitted. In addition stop-words were removed with the same provided German word list as described in the AMC preprocessing section. It is important to highlight that for further analysis the neutral classified items are converted to positive labeled items which results in a “non-negative” class.

[9] https://cst.ku.dk/sto_ordbase/
[10] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/
[11] https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tiger2dep/
[12] https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500

text	polarity
[(Top,), (qualifizieren, a), (Leute, n), (ver...	positive
[(Gott, n), (sein,), (danken,), (sein,), (f...	negative
[(",), (der,), (FPÖ, n), (werden,), (aus,)...	positive
[(Weil,), (ich, n), (mein,), (meist,), (Leu...	negative
[(Na,), (wer, n), (weis, n), (was, n), (da, r...	positive
...	...
[(Russland, n), (sein,), (in,), (wk1, n), (v...	positive
[(Was, n), (tendenziell, r), (kein,), (schlec...	positive
[(Was, n), (?,), (Unsinn, n), (!,), (der,),...	positive
[(wien, n), (verschrecken, v), (investoren, n)...	negative
[(Früher, r), (haben, v), (ich, n), (einen,),...	negative

Table 4.12: Standard posts after preprocessing

	word	WordNet-tag	short-pos-tag
0	fesch	ADJ	a
1	Zuckerl	NOUN	n
2	Topfenpalatschinke	NOUN	n
3	leiwand	ADJ	a
4	Ersparnis	NOUN	n
...	...	...	...
533	Schussattentat	NOUN	n
534	Exekution	NOUN	n
535	speiben	VERB	v
536	Brandleger	NOUN	n
537	Fotze	NOUN	n

Table 4.13: Overview of the Austriacisms data-set

4.5 Austriacisms

The list of Austriacisms shown in the table 4.13 was compiled by the project partners of University of Vienna based on the data-sets described in section 3.5. The manual assignment of the corresponding WordNet®-tags was performed together. Whereby the tags “ADJ”, “NOUN”, “VERB” and “ADV” were assigned to each word in the crafted list. Duplicates were manually removed from the data-set which resulted in a list of 539 words.

CHAPTER 5

Data Annotation

Chapter 5 describes the data annotation for the different data-sets in detail. The selection and preparation of the survey data and the evaluation of the results using statistical metrics was carried out as part of this thesis. The communication with the annotators over Prolific[1] as also the creation of the survey on SoSci Survey[2] was done by the project partners. Data annotation was required for the AMC data-set (see section 5.2) and for the list of Austriacisms (see section 5.3), neither of which have labels for the text-phrases/words. For each of the data-sets, the relevant steps for performing crowd sourcing, survey annotation, and the methods used are described. Different labeling approaches were used depending on the data-set.

5.1 Inter-Rater Reliability

The inter-rater reliability is an indicator which shows, how similar the annotators rated a list of items. Fleiss' kappa proposed by Fleiss [Fle74] is one of the metrics which is used to measure inter-rater reliability. This metric is an extension of Cohen's kappa. With Fleiss' kappa, inter-rater reliability is measurable without restriction by the number of annotators. The metric is explained in detail by Nichols et al. [NWCG10]. The Fleiss' kappa can be between "-1" and "+1". A value below zero indicates that the agreement is lower than it would be by chance, which is described in Landis & Koch. [LK77]. In their publication they propose an as they state "arbitrary" classification into categories which is shown in table 5.1. The resulting Fleiss' kappa depends on the number of classes and annotators.

[1] https://prolific.co/
[2] https://www.soscisurvey.de/

kappa statistic	strength of agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Table 5.1: Kappa statistic categorization proposed by Landis & Koch [LK77]

Table 5.1 shows the classification for two annotators and two classes. Furthermore, the domain in which the annotation was performed is important to decide whether the resulting Fleiss' kappa is "good" enough. The formula of this metric is defined as:

$$K = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} \tag{5.1}$$

- K = Fleiss' Kappa with a possible range of [-1,+1]
- $\overline{P}$ = Observed agreement
- $\overline{P_e}$ = Expected agreement

5.2 AMC

The AMC data-set described in section 4.3 is an integral part of this work. Labeling is very important for the later applied algorithms. Especially supervised approaches heavenly depend on a labeled data-set with high quality.

5.2.1 Survey Creation

Out of the 494.111 extracted text areas around Viennese politicians a total of 5.346 records were selected for the creation of the survey. The amount of labels was limited by the given financial budget. In total one internal pre-survey and two external surveys were conducted. The first survey contained 2.376 and the second 2.970 items. This results in a total of 5.346 items. The items for the survey were selected randomized over all news media, politicians and the entire time period with a fixed seed to allow reproducibility. It was ensured that politicians selected in the first survey did not appear in the second survey and vice versa. This allows further research by comparing the two conducted surveys. To ensure high quality responses, golden samples were introduced and internally checked for clarity. Overall, a given annotator had to achieve 75% correctness, otherwise the corresponding annotator was excluded from the results. Each item had to

be annotated at least three times to allow applying a majority vote. For edge cases where one label was positive, one negative and one neutral the item was labeled as neutral.

A total of 150 items had to be labeled by a particular annotator. Each 150 item-package contained 24 golden samples.

Golden Samples

From the extracted text areas, 24 golden samples were collected. It is important to mention that all golden samples were evaluated in a pre-survey by five German-speaking annotators to check the clarity of the statements. To avoid misinterpretation, some minor changes were made to express higher positive/negative sentiment. The following list shows a selection of the golden samples used during the crowd sourcing:

- "TR07_24 B1","Im Büro der zuständigen Stadträtin **Maria Vassilakou** (Grüne) begrüßt man es sehr, dass sich Bezirke für Radprojekte engagieren und verweist auf die stark wachsenden Zahlen. So waren im März 2017 am Opernring fast doppelt so viele Radfahrer unterwegs, wie im Vergleichszeitraum des Vorjahres. Konkret wurden pro Werktag 4635 Radfahrer gezählt.","5","0","0"
- "TR07_32 B1", "Alle kamen pünktlich, nur einer fehlte: **Werner Faymann**. Als der Bundeskanzler gestern um 15.30h beim Brüsseler Ratsgebäude vorfuhr, hatte der EU-Gipfel bereits ohne ihn begonnen. Auch das Treffen der Europäischen Sozialdemokraten versäumte er - Begründung: die Regierungserklärung im Bundesrat. Tags zuvor hatte Michael Spindelegger beim Treffen der Finanzminister gefehlt.","0","0","5"
- "TR07_34 B1", "Unrasiert und ungepflegt tritt heute FP-Chef **Strache** vor die TV-Zuschauer. Er wird vor allem zum skandalösen "Exil-Juden"-Sager befragt.","0","0","5","with modification: "unrasiert und ungepflegt""

The format per item is: "question ID, text, positive, neutral, negative, comment"; whereby the numbers indicate the amount of annotators which selected the corresponding class.

1st Survey

The first survey was conducted by labeling 2.376 items. In total a Fleiss' kappa of 0,295 was reached. The distribution was: 1.202 neutral, 598 positive, 576 negative; this indicates as expected for this domain a high amount of neutral items and class imbalance which needs to be taken into account for further processing. The Fleiss' kappa also suggests that the given textual phrases were not easy to label, and in particular the distinction between negative vs. neutral and positive vs. neutral were difficult to annotate.

2nd Survey

The second survey was conducted by labeling 2.970 items. The achieved Fleiss' kappa was 0,283. The distribution in the second survey was: 1.492 neutral, 787 positive, 691 negative; the distribution and the Fleiss' kappa is very similar to the first survey.

5.2.2 Crowd Sourcing

The pre-screening parameters needed to be set to a narrow range. Only persons which fulfill the following characteristics were approved:

- First language: German
- Nationality: Austria, German or Switzerland
- Current country of residence: Austria, Germany or Switzerland

This was necessary because the participants would not have been able to identify the positive, neutral or negative sentiment in the (Austrian) German texts.

At the time of conducting this crowd-sourcing step, using the split created by this work for the first and second annotation runs on SoSci[3] survey and Prolific[4], the project partners reported that a total of 3.000 active users met the jointly defined pre-screening parameters. A total of 182 participants took part in these surveys, with 24 not meeting the 75% threshold for the golden samples. This resulted in a total number of 158 annotators whose results were further used.

5.3 Austriacisms

In the AMC survey, labeling a representative portion based on phrases from the news media was shown to be challenging due to the neutral nature of news media articles. To improve the labeling process, Best-Worst Scaling (BWS) was used to assign labels to the words in the Austriacism list. The BWS approach used in this work was presented by Kiritchenko & Mohammad [KM17a, KM17b]. The methodology is based on the publication by Rouces et al. [RTBE18], who used the previously mentioned BWS method to create a Swedish sentiment lexicon.

[3] https://www.soscisurvey.de/
[4] https://www.prolific.co/

5.3.1 Methodology

The methodology applied by Rouces et al. [RTBE18] is described as:

$$l_{DA}(a,w) = \begin{cases} 1 & \text{if } a \text{ annotated } w \text{ as positive} \\ 0 & \text{if } a \text{ annotated } w \text{ as neutral} \\ -1 & \text{if } a \text{ annotated } w \text{ as negative} \end{cases} \tag{5.2}$$

$$sen_{DA}(w) = \frac{\sum_{a \in A_{DA}} l_{DA}(a,w)}{|A_{DA}|} \tag{5.3}$$

$$W_{BWS} = \{w : w \in W_{DA} \wedge |sen_{DA}(w)| \geq b\}. \tag{5.4}$$

- a = one annotator
- w = one word
- $sen_{DA}(w)$ = sentiment of one item
- A_{DA} = set of annotators
- $|A_{DA}|$ = number of annotators wich labelled the specific item w
- W_{DA} = set of words

The following formula shows which element can be marked as non-neutral. It is particularly important to note that in the referenced work three annotators were employed and therefore a majority vote was formed, resulting in a b value of 2/3:

$W_{BWS} = \{w : w \in W_{DA} \wedge |sen_{DA}(w)| \geq 2/3\}$

This formula was adjusted due to our varying number of annotators per item to:

$W_{BWS} = \{w : w \in W_{DA} \wedge |sen_{DA}(w)| \geq b\}$

whereby b is the required agreement over the specific item w which depends on the number of annotators who have labelled that specific item.

After this step they applied the BWS scaling algorithm which is based on the publication of Kiritchenko & Mohammad [KM17a, KM17b]. This approach is further explained in the Lexicon based approach chapter in the subsection 6.1.2.

5.3.2 Direct Annotation Survey

For the direct annotation survey, in addition to the classes “positive”, “neutral” and “negative”, a class “unknown” was added. This was required to ensure that annotators only tag items that they know. Internal testing has shown that it is difficult to know all regional words.

word
antisozial
Ausländerfeindlichkeit
denunzieren
...
stressfrei
umarmen
verzaubernd

Table 5.2: Direct annotation survey

Survey Creation

The creation of a set of golden samples was required to check the quality of the labeling done by the annotators of the crowd sourcing survey. The same 75% threshold was set as an approval rate. Further survey creation steps were not required that is why in this survey single words are labeled instead of text phrases.

Golden Samples

For the creation of the direct annotation survey in total 25 golden samples was used. A set of them is shown in table 5.2.

Results

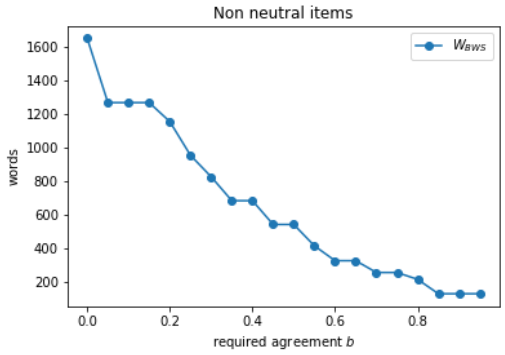

Figure 5.1: Direct annotation survey: comparison of different “b” settings

In total 1.600 words were labeled. It was ensured that each item was labeled at least three times. Before finding the non neutral items all unkown items and not answered ratings were excluded. The labels were converted to fit into the algorithm shown in the referenced paper (-1 = negative, 0 = neutral, +1 = positive).

After setting the in the methodology 5.3.1 mentioned b value to a value of 1/2 the 1.600 words were reduced to 544. This was done because the survey conducted collected between three and six annotations per word. By setting the b value to 1/2, a majority vote is formed in the minimum case of three annotations. In all other cases, at least two of all annotations agree. The influence of setting different thresholds for the extraction of the most positive and most negative words is shown in figure 5.1. The selected 544 words were the most positive and most negative words contained in the direct annotation survey.

A total of 18 annotators participated in the direct annotation survey. Only one annotator did not meet the requirements and was excluded. This leaves a total of 17 annotators.

5.3.3 Main Annotation Survey

The main survey was conducted by utilizing the in the work of Kiritchenko & Mohammad [KM17a, KM17b] BWS scaling method. The annotators were asked to label the best and worst out of a set of four items. A Perl script for the generation of the required tuples is available on the website[5] of Saif M. Mohammad. A scaling factor of two and four items per tuple was set as parameter, which resulted in a total of 4.417 tuples.

Survey Creation

The pre-screening parameters were set more strict than in the AMC crowd sourcing. This was required because, given that only Austrian German speaking people are able to correct label the sentiment of the given list of Austriacims. The parameters were set to:

- First language: German
- Nationality: Austria
- Current country of residence: Austria

Golden Samples

In total 20 golden samples were created for the main survey. In table 5.3 a part of them is shown. One interesting fact is that some of the annotators labeled the word “Zucker” which stands for “sugar” as “Bestitem” instead of labeling the word “angstfrei” which stands for “fearless” and was as “BestItem” intended. Therefore this golden sample was excluded for the calculation of the correctness. In general this shows the difficulty of performing crowd sourcing. Things tend to go wrong easily by selecting wrong golden samples.

positiv	neutral	neutral	negativ
angstfrei	Hose	Salz	antisozial
lohnenswert	Zucker	Wasser	Seuche
begeisterungsfähig	Suppe	Raum	Ausländerfeindlichkeit
...	...	...	...
entspannend	Butter	Jänner	denunzieren
Freundin	gehen	Kasten	einschläfern
herzerfrischend	Tag	gehen	Handlungsunfähigkeit

Table 5.3: Main survey golden samples

item1	item2	item3	item4	bestItem	worstItem
Rodel	Knödel-akademie	Keiler	Gelenks-beschwerden	Rodel	Gelenks-beschwerden
brennheiß	Storno-versicherung	Scherz(e)l	sich ausgehen	sich ausgehen	brennheiß
Steirer-anzug	Causa	Pönale	Lokal-augenschein	Lokal-augenschein	Steireranzug
Alumnat	Beiwagerl	Servus	kiefeln	Servus	kiefeln
Patchen-kino	Aufnahme-stopp	Straßen-erhalter	Marmeladinger	Straßen-erhalter	Aufnahme-stopp
...	...	...	...	...	...
ferten	Ermäßigungs-ausweis	Halbpreispass	versumpern	Ermäßigungs-ausweis	versumpern
Zuhaus	Bramburi	Mistbauer	Beiwagerl	Zuhaus	Mistbauer
Oja!	ludeln	Rettung	gar	Oja!	ludeln
Stützlehrer	Mascherl	Einspänner	grauslich	Mascherl	grauslich
Jausenbrot	enthaften	versperren	Schubhaft	Jausenbrot	Schubhaft

Table 5.4: Overview of the Austriacisms labeling result after performing the main survey

Results

The tuple generation was based on the 544 items of the direct annotation. 4.417 tuples were generated during the tuple generation. This tuples were labeled during the BWS process which is shown in table 5.4. In this table a set of the results based on the crowd sourcing is shown. The column "BestItem" refers to the best item which was chosen by the annotator. "WorstItem" refers to the worst item of all four different presented words.

In total 40 annotators participated in this main trial whereby 6 needed to be excluded from the results. They did not reached the golden sample threshold of 75%. A split-half reliability of 0,9159 (+/-0,0051) was reached with this approach. This indicates a very high reliability.

[5]https://saifmohammad.com/WebPages/BestWorst.html

CHAPTER 6

Lexicon Based Approach

Lexicon based approaches are widely used for performing sentiment analysis especially if the domain is narrow and the available labeled data are limited (subsection 2.3.2). This chapter describes the dictionary based approach used in this work as well as the created language resource which is called "Austrian Language Polarity in Newspapers (ALPIN)". Figure 6.1 describes the required datasets used in this approach. The approach is divided into two sections, using the AMC (section 6.2) and STANDARD (section 6.3) datasets by applying the SPLM algorithm (subsection 6.1.1) and the Austriacism list by applying the BWS algorithm (subsection 6.1.2). Furthermore, the two methods are combined in the post-processing step (subsection 6.6) and finally an evaluation of the results is performed (subsection 6.7).

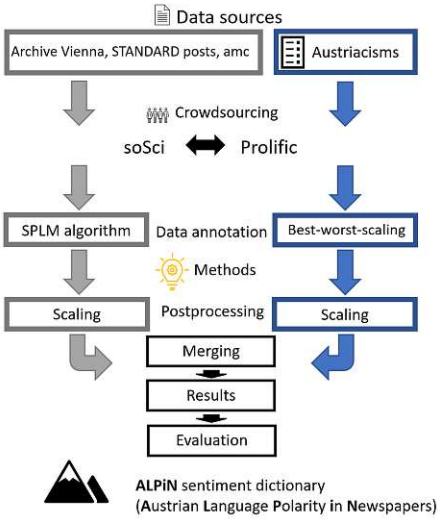

Figure 6.1: ALPIN - workflow

6.1 Methodology

Two different approaches are used for labeling the given data sets. In the next subsection 6.1.1 the SPLM algorithm is described. Then, in the subsection 6.1.2, the BWS algorithm is presented. Two different methods were used, since the SPLM algorithm is very cost efficient in terms of the amount of labels required during annotation and the BWS method in turn provides higher agreement between the evaluators, with the disadvantage that a higher number of labels are required.

6.1.1 SPLM Algorithm

There are several methods for computing sentiment scores of words given a labeled data-set [SD21]. One of the methods proposed by Sharma & Dutta [SD21] is called SPLM. This approach is used in this thesis to generate sentiment scores based on the labeled AMC data-set, as it performed well during evaluation in the referenced paper, especially in the "book" category. The SPLM algorithm proposed by Almantarneh & Gamallo [AG18] is defined and in their paper described by the following formulas:

$$RF_c(w) = \frac{freq(w,c)}{Total_c} \tag{6.1}$$

$$Avn(w) = \frac{\sum_{c=1}^{B} RF_c(w)}{B} \tag{6.2}$$

$$Avp(w) = \frac{\sum_{c=B+N_t}^{N} RF_c(w)}{B} \tag{6.3}$$

$$D(w) = Avp(w) - Avn(w) \tag{6.4}$$

- $RF_c(w)$ = relative frequency of a certain word in a specific rating label category
- $freq(w,c)$ = frequency of the corresponding word per rating label category
- w = word
- c = rating label category
- $Totalc$ = number of tokens in one rating label category
- $Avn(w)$ = average of the negative scores
- $Avp(w)$ = average of the positive scores
- $D(w)$ = resulting sentiment score (a value of 0 results in excluding the word from further usage)

This algorithm is shown as effective for the creation of a domain-specific sentiment lexicon in their publication.

	polarity	text
0	neutral	[(so, r), (eine,), (Ansinnen, n), (scheinen, ...
1	negative	[(Roland, n), (Sperk, n), (Vorsitzende, n), (d...
2	neutral	[(feiern, v), (werden, v), (in,), (Szenelokal...
3	neutral	[(in,), (die,), (ÖVP, n), (machen, v), (sich...
4	neutral	[(bei,), (eine,), (Erfolg, n), (die,), (Vol...
...	...	...
5310	negative	[(bei,), (ÖVP-Parteivorstand, n), (haben, v),...
5311	neutral	[(Zukunftswerkstätte, n), (I., n), (Schönlater...
5312	positive	[(die,), (gepflanzt, a), (Baum, n), (wollen, ...
5313	positive	[(in,), (Zug, n), (die,), (EU-Programm, n), ...
5314	neutral	[(jetzt, r), (laufen, v), (parallel, r), (eine...

Table 6.1: AMC labeled data-set (5.315 rows)

6.1.2 Best-Worst Scaling (BWS) Algorithm

For the generation of sentiment scores the method proposed by Kiritchenko & Mohammad [KM17a, KM17b] is utilized. BWS can be done by using different algorithms in their approach the method called "Counts Analysis" described by Orme [Orm09] is used. Counts analysis determines the percentage of positive and negative ratings of a word. At the end, the negative percentage is subtracted from the positive percentage. This results in a range of [-1,+1].

6.2 AMC Data-set

The following tables show the different steps in the preparation and application of the methods. Table 6.1 shows the labeled data-set that was used as input for the SPLM algorithm. Table 6.2 and 6.3 show the corresponding word list after applying the SPLM procedure. Even in this small subset, a reference to the Vienna area is recognizable.

6.3 STANDARD Data-set

The same approach by utilizing the SPLM algorithm is used for the STANDARD data-set. Table 6.4 shows the data-set after preprocessing. Table 6.5 and 6.6 show the output after applying SPLM.

6.4 AMC & STANDARD Data-set

In addition to the creation of individual sentiment dictionaries based on only AMC or STANDARD labeled data this step combines the labeled data-set of both sources. SPLM

	word	tag	D
2960	neu	a	2.339113e-03
4401	Wien	n	2.121626e-03
2929	Jahr	n	1.574550e-03
3785	Wiener	a	1.536313e-03
544	Michael	n	1.380177e-03
...	...	...	...
2272	gelten	v	3.097988e-06
6044	Sitzung	n	3.045782e-06
327	ziehen	v	5.220662e-08
2814	glauben	v	5.220662e-08
6245	Lösung	n	5.220662e-08

Table 6.2: AMC data-set after applying SPLM: positive words (2.368 rows)

	word	tag	D
6553	Peter	n	-0.001957
0	Westenthaler	n	-0.001913
1945	Pilz	n	-0.001766
4271	ÖVP	n	-0.001630
7076	kritisieren	v	-0.001574
...	...	...	...
2066	Fürsprecher	n	-0.000006
2024	Beste	n	-0.000006
7201	Schelling	n	-0.000006
4202	deutlich	r	-0.000006
5135	Fritz	n	-0.000006

Table 6.3: AMC data-set after applying SPLM: negative words (2.448 rows)

is applied on the combined data-set. Table 6.7 shows the input data-set whereby table 6.8 and 6.9 display the result after applying the algorithm.

6.5 Austriacisms

The Austriacism list shown in table 6.10 was generated by using the in subsection 6.1.2 outlined methodology.

	polarity	text
0	positive	[(Top,), (qualifizieren, a), (Leute, n), (ver...
1	negative	[(Gott, n), (sein,), (danken,), (sein,), (f...
2	positive	[(der,), (FPÖ, n), (werden,), (aus,), (alle...
3	negative	[(Weil,), (ich, n), (mein,), (meist,), (Leu...
4	positive	[(Na,), (wer, n), (weis, n), (was, n), (da, r...
...	...	...
3594	positive	[(Russland, n), (sein,), (in,), (wk1, n), (v...
3595	positive	[(Was, n), (tendenziell, r), (kein,), (schlec...
3596	positive	[(Was, n), (Unsinn, n), (der,), (Linguistik, ...
3597	negative	[(wien, n), (verschrecken, v), (investoren, n)...
3598	negative	[(Früher, r), (haben, v), (ich, n), (einen,),...

Table 6.4: STANDARD labeled data-set (3.599 rows)

	word	tag	D
4062	geben	v	0.002972
3255	Kind	n	0.001499
5064	Problem	n	0.001313
3928	Frau	n	0.001264
1114	Mann	n	0.001239
...	...	...	...
1175	Ausländer	n	0.000003
4222	beherrschen	v	0.000003
4337	weiss	n	0.000003
3625	leicht	r	0.000003
1305	Wahl	n	0.000003

Table 6.5: STANDARD data-set after applying SPLM: positive words (2.718 rows)

	word	tag	D
2224	Flüchtling	n	-2.387050e-03
1569	schon	r	-1.395659e-03
3606	ja	r	-1.157692e-03
3485	Land	n	-9.692610e-04
3916	Aktivist	n	-9.606781e-04
...	...	...	...
1001	feststellen	v	-3.573640e-06
2459	kaum	r	-4.785182e-07
4367	Beispiel	n	-4.785182e-07
1954	Chance	n	-2.392591e-07
1943	Zukunft	n	-2.392591e-07

Table 6.6: STANDARD data-set after applying SPLM: negative words (2.399 rows)

	polarity	text
0	neutral	[(Ansinnen, n), (scheinen, v), (Chef, n), (Bür...
1	negative	[(Roland, n), (Sperk, n), (Vorsitzende, n), (Ö...
2	neutral	[(feiern, v), (Szenelokal, n), (Stamperl, r), ...
3	neutral	[(ÖVP, n), (sogenannt, a), (Westachse, n), (En...
4	neutral	[(Erfolg, n), (Volksbegehren, n), (Jänner, n),...
...	...	...
8909	positive	[(Russland, n), (wk1, n), (vorzeirig, r), (aus...
8910	positive	[(tendenziell, r), (schlecht, a), (Tausch, n),...
8911	positive	[(Unsinn, n), (Linguistik, n), (turn, n), (bes...
8912	negative	[(wien, n), (verschrecken, v), (investoren, n)...
8913	negative	[(Früher, r), (vierteltelefon, n), (beantragen...

Table 6.7: AMC & STANDARD data-set combined (8.914 rows)

	word	tag	D
4966	geben	v	1.056874e-03
2298	Frau	n	1.028305e-03
6179	Jahr	n	9.792642e-04
2161	neu	a	9.567707e-04
8513	Mann	n	8.444807e-04
...	...	...	...
8457	unterbringen	v	1.124058e-06
10431	ÖVP-Generalsekretärin	n	1.124058e-06
2158	öffentlich	a	8.967702e-07
219	Mehrheit	n	8.967702e-07
5123	laufen	v	8.967702e-07

Table 6.8: AMC & STANDARD combined after applying SPLM positive words (4.605 rows)

	word	tag	D
6305	Flüchtling	n	-1.189029e-03
3076	Peter	n	-1.078103e-03
5595	ÖVP	n	-1.002944e-03
6622	Westenthaler	n	-9.941538e-04
141	Pilz	n	-9.196767e-04
...	...	...	...
918	feststellen	v	-1.351346e-06
540	Diskussion	n	-4.545758e-07
2191	Politiker	n	-4.545758e-07
10176	notwendig	r	-2.272879e-07
3940	klar	r	-2.272879e-07

Table 6.9: AMC & STANDARD data-set combined after applying SPLM: negative words (4.324 rows)

	word	score
0	fesch	0.882
1	Zuckerl	0.879
2	Topfenpalatschinke	0.857
3	leiwand	0.853
4	Ersparnis	0.844
...	...	...
532	Schussattentat	-0.844
533	Exekution	-0.848
534	speiben	-0.875
535	Brandleger	-0.879
536	Fotze	-0.969

Table 6.10: Austriacism list (537 rows)

6.6 Postprocessing

Due to the different approaches used, on the one hand the SPLM method for the AMC & STANDARD data-set and on the other hand BWS for the Austriacism list, a scaling of the resulting sentiment scores to a common range was necessary. Sentiment scores were scaled to a range from "-1" to "+1", with "-1" representing the most negative sentiment and "+1" representing the most positive sentiment. The MaxAbsScaler of the Python package Scikit published by Pedregosa et al. [PVG+11] is utilized. With the help of this scaler algorithm it is possible to scale a distribution to a certain range without changing the relation and relative distance between the sentiment scores.

6.6.1 Scaling: AMC & STANDARD Data-set

Figure 6.2 shows that the SPLM algorithm produces very small numbers whereby they are clustered around zero. After scaling the distribution is mapped to the range [-1,+1]. The reason behind the clustering around zero is that the SPLM algorithm calculates the relative frequencies which leads to small numbers if the occurrence of the word in relation to the tokens in the corresponding class is low.

6.6.2 Scaling: Austriacism List

In figure 6.3 it is shown that the sentiment scores of the Austriacism list are well distributed. Nevertheless to align this list to a common range the same scaling procedure was applied.

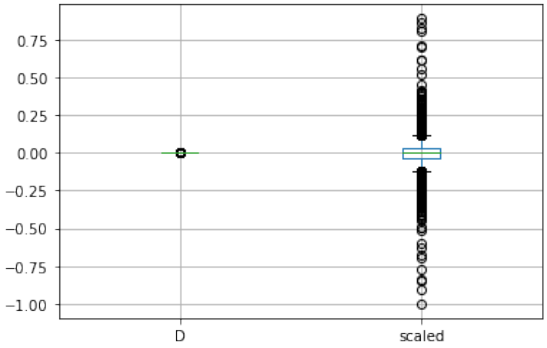

Figure 6.2: AMC & STANDARD data-set scaling - comparison

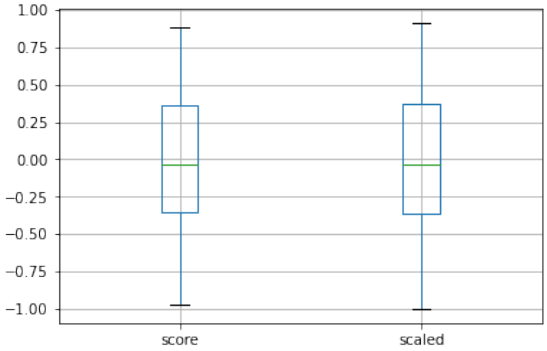

Figure 6.3: Austriacism list scaling - comparison

6.6.3 Merging AMC & STANDARD Data-set with the Austriacism List

After matching the different data-sets in the subsections 6.6.1 & 6.6.2, the data-sets were merged. If a word was present in both data-sets, the word was omitted from the AMC & STANDARD data-set. This is due to the fact that individual words were flagged in the Austriacism-list, making the accuracy of the sentiment scores of these individual words higher than that of the sentiment scores automatically generated based on SPLM. In particular, these word intersections are important for the evaluation and comparison of the dictionaries, which will be done in the next section. The final data-set consists of 9.435 rows and was created by merging the AMC & STANDARD data-set with the Austriacism list. This data-set is shown in table 6.11. The combined version of this data-set is named “Austrian Language Polarity in Newspapers (ALPIN)”.

	word	short-tag	scaled
0	fesch	a	0.910
1	Zuckerl	n	0.907
2	geben	v	0.889
3	Topfenpalatschinke	n	0.884
4	leiwand	a	0.880
...	...	...	...
9430	speiben	v	-0.903
9431	Peter	n	-0.907
9432	Brandleger	n	-0.907
9433	Fotze	n	-1.000
9434	Flüchtling	n	-1.000

Table 6.11: Combined data-set with Sentiment scores based on SPLM & BWS

6.7 Evaluation

Evaluation was performed by contrasting the dictionary which is based on AMC & STANDARD data-set and the Austriacism list against the labeled STANDARD and AMC data-sets. In addition an external evaluation was performed. It is important to mention that at this point of time there is no Austrian German data-set in the area of news-media and politics to evaluate the ALPIN dictionary against.

6.7.1 Methodology

To evaluate the ALPIN dictionary against the labeled STANDARD data-set and the AMC data-set, feature engineering was required. The same feature engineering and evaluation methodology as in the paper of Almatarneh & Gamallo [AG18] was applied. Feature engineering was done by calculating how many words of a given text item are positive and how many negative. By looking up the words and their sentiment in the ALPIN dictionary. Additionally the proportion of positive and negative text areas is used as a third feature. After calculating the features they were used to train a simple support vector machine (SVM) with a linear kernel [PVG+11]. KFold (5-Fold) was applied to ensure that the calculated metrics are representative for the given data-set. The metrics used for evaluating the results were: accuracy, precision, recall and f1 score.

6.7.2 External Evaluation Data-sets

Two external data-sets were used for performing the evaluation. First the “German Polarity Clues (GPL)” is introduced. After GPL a brief introduction of the second language resource is given which is named “Affective Norms (AN)”.

	lemma	wordform	classification
0	Abfangschirm	NN	positive
1	Abgeklärtheit	NN	positive
2	Abgeschlossenheit	NN	positive
3	Abgleich	NN	positive
4	Abgott	NN	positive
...	...	...	...
5924	übertreiben	VV	negative
5925	übertreten	VV	negative
5926	übertrieben	AD	negative
5927	übervorteilen	VV	negative
5928	überwältigen	VV	negative

Table 6.12: German Polarity Clues (9.561 rows)

German Polarity Clues (GPL)

German Polarity Clues (GPL) is a publicly available German sentiment dictionary. GPL was, introduced and created by Waltinger [Wal10] and provides a list of words with sentiment values and respective word form. A semi-automatic translation approach was used for this language resource. The goal of the research by Waltinger [Wal10] was to create a new German dictionary based on existing English dictionaries. Table 6.12 shows an outline of the in total 9.561 words contained in the sentiment dictionary.

Affective Norms (AN)

Affective Norms (AN) is a sentiment resource introduced by Köper & Schulte im Walde [KSiW16]. This resource is a collection of German sentiment dictionaries and translated English dictionaries combined together. They describe in their work that a supervised algorithm was used to automatically calculate the scores for each rating type. The rating types of this resource are: abstractness, arousal, imageability and valence. In this work, only valence is used. For the purposes of this work, the rating scale was adjusted to "-1" for all values below five and "+1" for all values above five. All words with a score of exactly five were excluded. The result of the preprocessing is shown in table 6.13.

6.7.3 Results

In this comparison the sentiment dictionaries GPL, AN and ALPIN are used. The feature calculation is performed as described in the methodology in subsection 6.7.1. It is important to mention that the ALPIN dictionary is based on the labeled data of the STANDARD data-set and the AMC data-set labeled data which is shown in table 6.7. As a result there is a dependency between the the data-set during evaluation and the created ALPIN sentiment dictionary.

	lemma	classification
0	sein	1
1	in	1
2	ein	1
3	werden	1
4	von	1
...	...	...
351606	bebauter	-1
351608	Trauermett	-1
351610	traditionslinke	-1
351615	Täterakte	-1
351616	deutschstämmiger	-1

Table 6.13: Affective Norms (351.502 rows)

dictionary	accuracy	precision	recall	f1 score
GPL	0.526	0.528	0.986	0.688
AN	0.530	0.530	**1.0**	0.693
ALPIN	**0.768**	**0.778**	0.794	**0.783**

Table 6.14: ALPIN against STANDARD data-set

Sentiment Dictionary Against STANDARD Data-set

In table 6.14 it is shown that GPL and AN were not able to correctly identify a positive/negative sentiment score. In contrast, ALPIN was able to correctly predict the labels in most of the cases.

Sentiment Dictionary Against AMC Data-set

In this comparison in table 6.15 the results are much better than in the first one with the STANDARD data-set. GPL and AN achieve good results, nevertheless the results of ALPIN dictionary are better. Comparing the results of the last evaluation with the current evaluation, there is an improvement especially in the metrics of GPL and AN.

This suggests that the early decision to combine the neutral and positive classes of the labeled STANDARD data-set into one “positive class” may not have been as effective as originally expected.

Evaluation with Train-test Split

To overcome the dependency between the ALPIN sentiment dictionary and the STANDARD & AMC data-sets, an alternative evaluation strategy was performed. For this

dictionary	accuracy	precision	recall	f1 score
GPL	0.647	0.641	0.738	0.686
AN	0.657	0.660	0.660	0.670
ALPIN	**0.822**	**0.828**	**0.840**	**0.830**

Table 6.15: ALPIN against AMC data-set

ratio	accuracy	precision	recall	f1 score
0.6	0.542	0.543	0.783	0.633
0.7	0.549	0.563	0.626	0.577
0.8	0.567	0.557	**0.912**	0.692
0.9	**0.614**	**0.596**	0.831	**0.693**

Table 6.16: Results STANDARD data-set (train-test)

purpose, the labeled STANDARD & AMC data-set was split into a training and a test data-set before applying the SPLM algorithm. This allows the creation of a sentiment dictionary based on a portion of the labeled data-set and the ability to use a "real" unused test data-set to validate the results. This process was performed with different ratios to show how the size of the training data-set affects the performance of the language resource created. This is particularly important because the amount of labeled data is small. Apart from the division into training and test data, similar processing was performed before applying the SPLM algorithm. For feature calculation, the number of positive and negative words and the ratio of positive to negative were calculated. The features were used to perform K-fold with five folds. Prediction was performed using a linear SVM kernel and results were calculated using cross-validation. Metrics for each fold were averaged, resulting in a set of accuracy, precision, recognition score and f1 score for each train-test split.

The results are shown in table 6.16, 6.17 and 6.18. They are very similar to the last evaluation performed. It is shown that it is possible to measure sentiment scores based on the collected labeled data. In addition, this comparison was performed by validation with an external data set which proofs that the prediction of external data-sets in the domain of news media an politics, in the Austrian German area, is possible.

Comparison of Word Intersections

Another comparison was done by having a look at words which occurred in the AMC & STANDARD posts data-set and the Austriacism list. Table 6.19 shows the top ten and bottom ten words which were contained in both data-sets. In total there was an overall intersection of 32 words. From the 21 positive words of the Austriacism list 16 were also positive in the AMC & STANDARD data-set. Out of 11 negative words of the Austriacism list only one was not negatively annotated in the AMC & STANDARD

ratio	accuracy	precision	recall	f1 score
0.6	0.644	0.655	**0.716**	0.682
0.7	0.647	0.670	0.689	0.671
0.8	0.678	0.726	0.643	0.675
0.9	**0.704**	**0.739**	0.702	**0.719**

Table 6.17: Results AMC data-set (train-test)

ratio	accuracy	precision	recall	f1 score
0.6	0.588	0.586	**0.777**	0.664
0.7	0.607	0.609	0.727	0.659
0.8	0.577	0.583	0.707	0.634
0.9	**0.615**	**0.618**	0.774	**0.671**

Table 6.18: Results STANDARD & AMC data-set (train-test)

posts data-set. This shows that the results of the different algorithms SPLM & BWS resulted in similar results whereby it also indicates that the negative labeled part of the data-set is more accurate as the positive side.

word	Austriacism list (sentiment score)	AMC & STANDARD data-set (sentiment score)
Wiese	0.750	0.027
Karenz	0.742	0.040
Angelobung	0.729	-0.051
Ehrenzeichen	0.710	0.067
Gehalt	0.645	0.211
aufrecht	0.625	-0.031
maturieren	0.625	0.027
ÖAMTC	0.562	-0.031
einbringen	0.548	-0.123
Team	0.516	0.166
...	...	...
klagen	-0.312	-0.054
angreifen	-0.344	-0.005
Fleck	-0.376	-0.031
Einvernahme	-0.387	-0.031
Freunderlwirtschaft	-0.438	-0.031
versperren	-0.486	-0.031
Mist	-0.594	-0.031
sekkieren	-0.688	-0.031
exekutieren	-0.838	-0.004
Exekution	-0.875	-0.027

Table 6.19: Top and bottom 10 word intersections: Austriacisms vs. AMC & STANDARD data-set (32 rows in total)

CHAPTER 7

Supervised Approaches

In addition to lexicon based approaches, supervised approaches are one major research topic in performing sentiment analysis. This chapter gives a overview about the different methods used in this area. In section 7.1 the required tools for performing the analysis are listed. Section 7.2 outlines the different methods used for performing the supervised approaches. It is important to mention that this section is split into a baseline section were more traditional approaches are explained and dedicated sections for the state-of-the-art algorithms. In the last section 7.3 the results and evaluation are presented.

7.1 Tooling

For the various supervised approaches different data-science related packages of the Python programming language were used. One of the most important packages is Scikit-learn by Pedregosa et al. [PVG+11]. In addition the NLTK package by Bird et al. [BKL09] was utilized for NLP processing task and for retrieval of the German stopword list. A dedicated fastText implementation by Joulin et al. [JGBM16] and the Transformers library by Wolf et al. [WDS+20] was utilized in the state-of-the art implementations.

7.2 Methodology

The different baseline algorithms are explained in subsection 7.2.1 . After outlining the baseline a more sophisticated algorithm called fastText, is presented in subsection 7.2.2. The current state-of-the art implementation is shown in subsection 7.2.3 whereby in this one two different variations of the model are outlined. It is important to mention, that the neutral class is excluded and binary (positive, negative) classification is performed during the supervised approaches. This was done to keep in track with the binary classification used in the lexicon-based approach.

7.2.1 Baseline

DummyClassifier

Before implementing the baseline and more complex algorithms, it is beneficial to look at how well a random classifier performs. Therefore the DummyClassifier feature of the Scikit-learn framework was utilized. In the thesis four different variations were used and are shortly described a :

- DummyClassifier v1 ("stratified"): returns predictions based on a multinomial distribution which is parametrized by the empirical class prior probabilities.
- DummyClassifier v2 ("most_frequent"): returns most frequent class label based on "y" Whereby "y" is the observed variable.
- DummyClassifier v3 ("uniform"): returns uniformly generated predictions based on "y".

Remark: The description of the different classifiers is based on the Scikit-learn framework description[1].

PassiveAggresiveClassifier

This classifier[2] is based on the publication of Crammer et al. [CDK+06]. The PassiveAggresiveClassifier allows binary classification and belongs to the category "Online machine learning" (= process data in a sequencial row). The naming of the classifier is given due to being passive on correct predictions and aggressive learning behaviour by wrong predictions.

RidgeClassifierCV

RidgeClassifierCV[3] is able to perform multi class preditions. The classifier is based on the Ridge regression which is explained in detail in the work of McDonald [McD09].

LogisticRegressionCV

LogisticRegressionCV[4] is based on logistic regression. There are a number of different solvers that can be used for the optimization problem. It is possible to solve single class also as multiple class problems. In this thesis the "saga" solver was used which is known to be faster as the other solvers for bigger data-sets. The "saga" solver is further described in the work of Defazio et al. [DBLJ14].

[1] https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.PassiveAggressiveClassifier.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifierCV.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html

SGDClassifier

SGDClassifier[5] is based on a regularized linear models combined with stochastic gradient descent. This classifier is also supporting multi class predictions. As loss function in this work "log" is used which referes to logistic regression.

SVC

SVC[6] stands for "Support Vector Classification" and supports multi class labeling. Many different kernels are supported ("linear", "poly", "rbf", ...) whereby in this work a "linear" kernel is used. The implementation is based on LIBSVM [CL11].

MLPClassifier

MLPClassifier[7] is a neural network which is supporting multi class labeling. MLP is defined as "Multi-layer Perceptron". Different solvers are possible whereby in this work the "adam" solver is used which is based on a stochastic gradient descent optimization. This classifier is further described in the work of Glorot & Bengio [GB10].

7.2.2 FastText

FastText is a framework proposed by Facebook Inc. and published by Joulin et al. [JGBM16]. They describe fastText as efficient text classification algorithm which is able to achieve good performance on a CPU based environment without the requirement of having expensive GPU resources. This framework is available as Python package. The project itself is compiled in C++. It is possible to perform different tasks e.g. learning word representation models and text classification models. This framework supports multi class labeling whereby for predicted each of the predicted classes the corresponding metrics (precision, recall and f1 score) are calculated.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

7.2.3 BERT

BERT stands for "Bidirectional Encoder Representations from Transformers" which was proposed by Devlin et al. [DCLT19] is often used in recent research to perform sentiment analysis tasks. In this section two different approaches are shown. There are many different BERT models trained on different languages and for specific domains. Neverthless there is currently no special BERT model trained for performing sentiment analysis in the area of politicial news in the Austrian or German area.

In this thesis two existing BERT models are used and fine-tuned by using the labeled AMC data-set collected with the help of crowd sourcing. The first model is based on a lightweight version of the original BERT model. The second variant is based on a crafted BERT model of German texts and news.

BERT (distilbert-base-german-cased)

DistilBERT[8] is a lightweight version of the original BERT model. This reduced model was introduced by Sanh et al. [SDCW20]. They state that this model is smaller, faster and keeps nearly the full language understanding capabilities with only a loss of 3% compared with the original BERT model.

BERT (dbmdz/bert-base-german-cased)

The dbmdz German Bert base cased model[9] is a BERT model crafted by the Bavarian State Library. This model is trained on different data-sets e.g. Wikipedia, the EU Bookshop corpus, news data, websites and many more. Especially valuable for this work is that in this model German news are included.

[8] https://huggingface.co/distilbert-base-german-cased
[9] https://huggingface.co/dbmdz/bert-base-german-cased

	polarity	text without stopwords
0	0	Roland Sperk Vorsitzende Österreichische Gewer...
6	0	Menschenrecht Selbstverständlichkeit sagen Bun...
11	1	wenig nehmen Werner Faymann wichtig Kampf Juge...
13	1	Degot sechst Intendantin Festival heuer -Jahr-...
17	0	Großteil Eigentum Stadt Wien stehend Teerag-As...
...	...	...
5306	0	berichten Weekend-Treff gut Klima erneut Durch...
5307	0	FP-LAbg. Gerhard Zeihsel fordern zumindest gem...
5309	0	ÖVP-Parteivorstand oberösterreichisch Landesha...
5311	1	gepflanzt Baum Spitzenkandidat Johannes Voggen...
5312	1	Zug EU-Programm Urban Wien Gürtel Lauf Jahr Gr...

Table 7.1: AMC data-set after preprocessing and filtering out neutral items (2.638 rows in total)

7.3 Results & Evaluation

7.3.1 Baseline

Calculation of the baseline performance was done by developing a Python script which allows to easily add and remove different machine learning models. Supported are models which are available in the Scikit-learn framework by having a common wrapper around these. The in subsection 7.2.1 mentioned models were applied on the given AMC data-set.

The AMC data-set was split into train & test data-set whereby the test size was 25% of the data-set. For each model cross-validation was performed on the remaining training data-set of the last step by using 5-folds and a scorer for the metrics “accuracy”, “precision”, “recall” and “f1 score”. With the help of the “classification_report” function the metrics per model were collected. For the feature engineering TF-IDF was utilized. With the setting min_df = 5 terms with a document freequency lower than five are excluded. Sublinear term frequency is set to true which changes the term “frequency only” to the “logarithm of the term frequency”. As norm “l2” is used which indicates that each output row vector uses the Euclidean distance. Finally the ngram_range is set to (1,2) which sets the lower and upper border for different n-grams in this case only unigrams and bigrams. As input data-source the AMC data-set is used. The preprocessed variant is utilized in which the stopword removal based on the lemmas for each text item was already performed. A outline of the data-set is shown in table 7.1.

The results of the baseline approaches based on binary classification are shown in table 7.2. The given metrics were calculated for both of the classes whereby in the shown table the average over both classes per metric was built. The best performing algorithms of the baseline are the PassiveAggresiveClassifier and the RidgeClassifierCV.

model name	accuracy	precision	recall	f1 score
DummyClassifier v1	0.515	0.542	0.507	0.524
DummyClassifier v2	0.526	0.526	**1.000**	**0.689**
DummyClassifier v3	0.521	0.543	0.571	0.556
PassiveAggressiveClassifier	**0.629**	**0.644**	0.657	0.650
RidgeClassifierCV	0.623	0.615	**0.755**	**0.678**
LogisticRegressionCV	0.614	0.614	0.723	0.663
SGDClassifier	0.603	0.621	0.625	0.623
SVC	0.623	0.637	0.660	0.648
MLPClassifier	0.609	0.624	0.648	0.635

Table 7.2: Baseline results of the test data-set (= average over all folds)

	fasttext format
0	___label___negative roland sperk vorsitzender de...
6	___label___negative menschenrechte seien keine s...
11	___label___positive wenig nimmt werner faymann w...
13	___label___positive degot wird 2018 als sechste ...
17	___label___negative die groteils im eigentum der...
...	...
5306	___label___negative wie berichtet hatten die wee...
5307	___label___negative fp labg gerhard zeihsel ford...
5309	___label___negative beim vp parteivorstand hat d...
5311	___label___positive den gepflanzten baum will sp...
5312	___label___positive im zuge des eu programms urb...

Table 7.3: AMC data-set after preprocessing and filtering out neutral items (2.638 rows in total)

7.3.2 FastText

FastText described in the subsection 7.2.2 is a more enhanced algorithm compared with the approaches of the baseline. In addition fastText requires a special labeling during preprocessing before the method can be applied. The method requires "___label___negative" and "___label___positive" before each item to indicate if the item is positive or negative. The preprocessed input data-set is shown in table 7.3.

A to the baseline similar separation of the data-set into train, test and validation data-set was performed. The data-set shown in table 7.3 was split into a train, test and validation data-set. Whereby 80% are assigned to the train data-set and 10% for each, the test and validation data-set. Additional preprocessing was performed as suggested by the

data-set	label	precision	recall	f1 score
validation	negative	0.670	0.627	0.648
	positive	0.678	**0.717**	0.697
test	negative	0.598	0.676	0.635
	positive	**0.754**	0.686	**0.718**

Table 7.4: FastText results split by label

fastText documentation[10] to optimize the result of the classifier which included removing punctuation, uppercase to lowercase, usage of word n-grams and encoding as UTF-8.

After evaluating different hyperparameter settings and additionally using the automatic hyperparameter optimization the final model parameters were set to: learning-rate="0.15", epoch="25', wordNgrams="2" and dim="100" (=size of word vectors). The results are shown in table 7.4. An increase over all metrics compared with the baseline is visible. Nevertheless the increase compared with the RidgeClassifierCV of the baseline is not so high.

7.3.3 BERT

BERT is currently one of the most advanced algorithms in the field of sentiment analysis. As described in subsection 7.2.3, two different models were used in this thesis. The same procedure was used for both of the pre-built BERT models which allows testing different models in a structured way. The Transformers library by Wolf et al. [WDS+20] was used as described in the tooling section for fine-tuning the two BERT models with the labeled AMC data-set. In addition to the already mentioned frameworks PyTorch by Paszke et al. [PGM+19] is utilized. PyTorch is an open source machine learning framework.

The labeled AMC data-set was split into a training (80%) and test (20%) data-set. After this step the corresponding BERT models "distilbert-base-german-cased" and "distilbert-base-german-cased" were retrieved from the HuggingFace[11] repository. This site is a platform for state-of-the-art machine learning models and frameworks. Tokenization was performed whereby the "max_length" of 512 tokens was set for the padding parameter as also truncation is activated. The Trainer API[12] was used for performing the fine-tuning process. The following parameters were set: evaluation_strategy="steps", eval_steps="250", per_device_train_batch_size="8", per_device_eval_batch_size="8", num_train_epochs="6", seed="42" and load_best_model_at_end="TRUE".

The results of the two different models after fine-tuning are shown in table 7.5 for the BERT (distilbert-base-german-cased) fine-tuned with the AMC data-set and in table 7.6 for the BERT (dbmdz/bert-base-german-cased) fine-tuned on AMC data-set. A

[10] https://fasttext.cc/docs/en/supervised-tutorial.html#preprocessing-the-data
[11] https://huggingface.co/
[12] https://huggingface.co/docs/transformers/master/en/main_classes/trainer#transformers.Trainer

step	training loss	validation loss	accuracy	precision	recall	f1 score
250	No log	0.450	0.794	0.822	0.771	0.796
500	0.430	0.574	**0.796**	**0.830**	0.764	0.796
750	0.430	1.001	0.771	0.745	0.851	0.795
1000	0.156	1.346	0.752	0.711	**0.884**	0.788
1250	0.156	1.283	0.790	0.783	0.826	**0.804**

Table 7.5: BERT (distilbert-base-german-cased) fine-tuned with the AMC data-set results by step

step	training loss	validation loss	accuracy	precision	recall	f1 score
250	No log	0.469	0.788	0.794	0.800	0.797
500	0.496	0.496	**0.796**	**0.795**	0.818	0.807
750	0.496	0.925	0.771	0.728	**0.895**	0.803
1000	0.214	1.349	0.778	0.738	0.891	0.807
1250	0.214	1.353	0.792	0.756	0.887	**0.816**

Table 7.6: BERT (dbmdz/bert-base-german-cased) fine-tuned with the AMC data-set results by step

model	accuracy	f1 score	loss	precision	recall
DBGC	0.796	0.796	0.574	**0.830**	0.764
DBMDZ	**0.800**	**0.801**	0.496	0.795	**0.818**

Table 7.7: Comparison “dbmdz/bert-base-german-cased” (DBMDZ) and “distilbert-base-german-cased” fine-tuned (DBGC)

comparison of the by the “load_best_model_at_end” function selected method is shown in table 7.7. After fine-tuning both BERT models they achieved similar performance metrics. The DBMDZ based model is slightly better than the competitor DBGC. Overall both BERT based models outperform all the other applied algorithms including the fastText approach.

CHAPTER 8

Discussion

In the lexicon based approach in chapter 6 and the supervised approaches in chapter 7 the effectiveness of models for predicting sentiment in this narrow domain is explained in detail. This chapter presents a general perspective and discussion on the results of this work. In the first section 8.1 the preprocessing and data annotation steps are discussed. This is followed by a comparison of the lexicon based approach and supervised approach in section 8.2. After that the distribution of politicians is discussed in section 8.3. The web application is shown in section 8.4 whereby a qualitative analysis based on the results of this work is performed. The web tool and the distributions shown in this section can be used to answer the media biases mentioned by Eberl et al. [EBW17]. In the end ethical questions are discussed in section 8.5.

8.1 Preprocessing & Data Annotation

Before data extraction and data annotation can begin, it is necessary to clearly define what the outcome of an analysis should be. Questions such as "How is bias defined?" must be raised and discussed in depth. This question is especially important if the labeling is done by external annotators. The annotators need to clearly understand what the definition of "bias" in this work is. Only if they understand that, they are able to label the data-set in a way so it is able to solve the given task. The concept of "bias" in this thesis is closely related to sentiment scores of politicians. Therefore, a common understanding of what should be labeled as positive, neutral, or negative is very important. Another important fact is that labeling itself is an highly emotional task. If an annotator is orientated on the political left wing the annotator might tend to label phrases of right winged politicians in a negative way which could also be related to an unconscious bias [TL16]. The crowd-sourcing itself was very successful, ensuring that payment was fair and that the survey was conducted on a platform from a European country to ensure

high standards in terms of data privacy and the General Data Protection Regulation (GDPR).

In the preprocessing section, it was shown that phrase-based extraction is effective in extracting text phrases that have a limited number of tokens due to legal requirements. Nevertheless with a higher financial budget labeling of more text phrases and applying aspect-based sentiment analysis as described in Do et al. [DPMA19] is recommended. During data annotation the need for golden samples became apparent very quickly. It was expected that labeling this type of data in the AMC crowd-sourcing lead to an inter-rater agreement which is not as high as it would be in other domains. Some of the annotators attempted to label arbitrarily and were successfully detected by the embedded golden sample based quality control. The two surveys conducted, which were required to label the Austriacism list, were very successful. The split-half reliability calculated after the main survey was exceptionally good, indicating that this approach fits very well within the scope of this work. In comparison, the AMC survey produced more moderate results, which can be explained by the neutral nature of news media. News media are not as easy to label with respect to sentiment scores as, for example, tweets or movie reviews.

8.2 Lexicon Based Approach Vs. Supervised Approaches

In this work, several different methods were applied to the given data-sets. It has been shown that both the lexical based approach and the machine learning approaches lead to good results. Looking only at the metrics, there is a clear winner. The BERT based approaches outperform all other machine learning and lexical approaches conducted. Nevertheless one needs to tackle that result also from a different perspective. As we know many lexical based approaches are explainable and transparent. It is relatively easy to verify why the sentiment score of a certain politician is e.g. positive and not negative. Analysing what the best performing algorithm in this thesis, BERT, is doing is not possible with the current state-of-the-art. Explainability, especially for deep learning algorithms, is currently under intense research and discussed in the scientific community. The lack of explainability is the reason why the model used in the web application is based on the lexicon based approach rather than the BERT based fine-tuned model.

8.3 Distribution of Politicians

For further analysis the frequency of the politicians over time and different media was analysed as described in the statistics subsection 4.3.6 of the AMC preprocessing part of this thesis. It has been shown that the Viennese news media landscape is strongly focused around a small set of certain politicians which is shown in figure 8.1.

Figure 8.2 shows the top fifteen politicians with the highest number of hits across all news media. A “hit” is defined as the occurrence of a politician in a paragraph. This graph clearly shows that the distribution of politicians follows a power law distribution. In particular, “Werner Faymann” and “Sebastian Kurz” are very well represented in

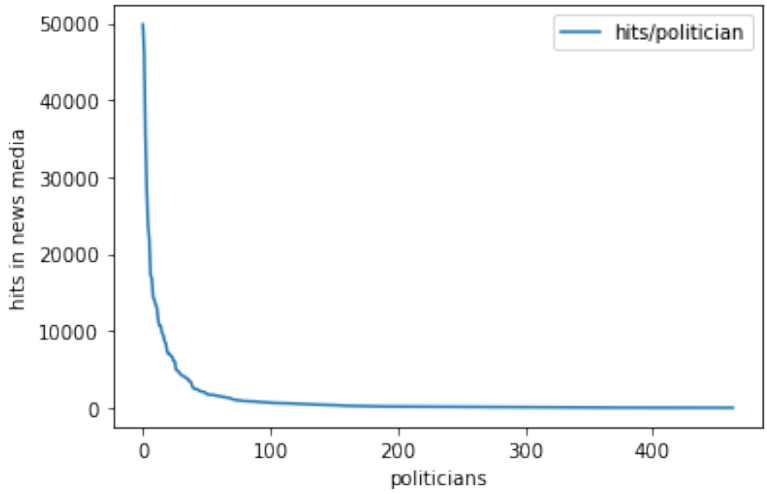

Figure 8.1: Politician distribution in the overall news media landscape

the news during the observed time-range. In addition the distribution over the different political parties is interesting. Seven out of fifteen politicians belong to the "Social Democratic Party of Austria (SPÖ)," a party that tends to the left on the political scale. Three politicians are related to the "Austrian People's Party (ÖVP)" which is liberal conservative and another three to "The Greens – The Green Alternative (Die Grünen)". "Die Grünen" is located on the centre left to left on the political scale. Only two politicians are from the "Freedom Party of Austria (FPÖ)" which is a party on the right to far right wing. The high amount of politicians related to the party "SPÖ" could be related to the focus around Viennese politicians and the long history of left wing politics. Since 1945 the mayors have been appointed by the "SPÖ".

Table 8.1 shows the top 15 politicians by their absolute values. In this table the fast decrease in the overall hit count over all news media is shown.

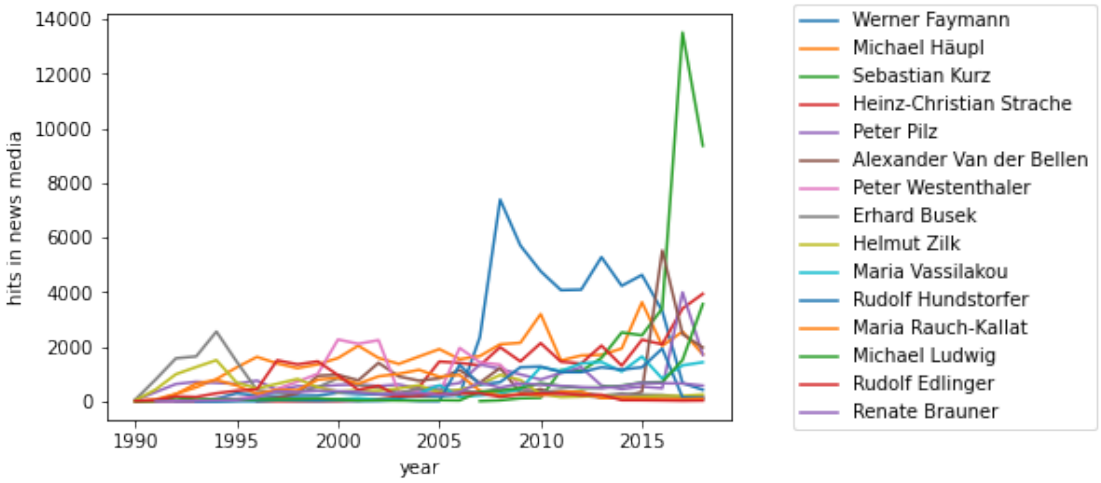

Figure 8.2: Top 15 politicians over all news media

	politician	hits	party
0	Werner Faymann	49833	SPÖ
1	Michael Häupl	46031	SPÖ
2	Sebastian Kurz	35349	ÖVP
3	Heinz-Christian Strache	28474	FPÖ
4	Peter Pilz	23737	Die Grünen / Liste Peter Pilz
5	Alexander Van der Bellen	21966	Die Grünen / President
6	Peter Westenthaler	17281	FPÖ/BZÖ
7	Erhard Busek	16758	ÖVP
8	Helmut Zilk	14427	SPÖ
9	Maria Vassilakou	14037	Die Grünen
10	Rudolf Hundstorfer	13421	SPÖ
11	Maria Rauch-Kallat	12989	ÖVP
12	Michael Ludwig	11422	SPÖ
13	Rudolf Edlinger	10654	SPÖ
14	Renate Brauner	10611	SPÖ

Table 8.1: Top 15 politicians over all news media

8.4 Web Application

The results of this thesis are used in a web application[1] which was created by the team at ACDH-CH. Qualitative evaluation is possible by selecting different politicians on the application.

[1] https://dysen-tool.acdh-dev.oeaw.ac.at/

Figure 8.3 shows the normalized frequency distribution over time for the most frequently occurring politician, Werner Faymann. The second figure 8.4 shows the distribution of sentiment over time. The two figures clearly show the beginning and end of Werner Faymann's political career peak. Also, once the normalized frequency increases to a higher level, a trend in sentiment scores over time is visible. The higher variance in the first years from 1994 to 2006 is related to the small amount of sentiment predictions in this period. A negative trend of the sentiment scores is visible.

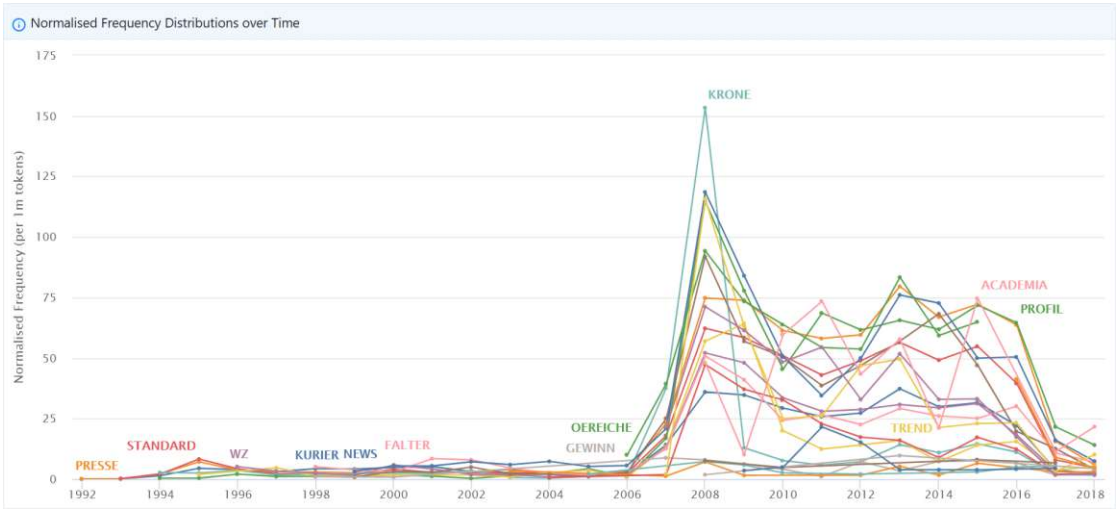

Figure 8.3: Normalised frequency distributions over time of Werner Faymann (Link: https://dysen-tool.acdh-dev.oeaw.ac.at/)

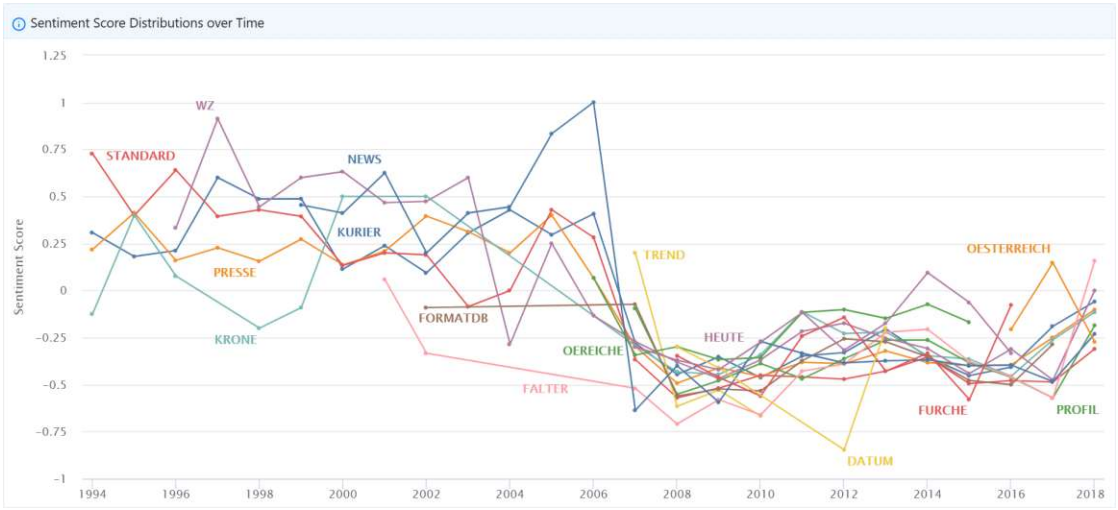

Figure 8.4: Sentiment Score Distributions over time of Werner Faymann (Link:https://dysen-tool.acdh-dev.oeaw.ac.at/)

With the results of this thesis, the normalised frequendcy distribution and sentiment scores over time for each of the news media, further qualitative research is possible. For example one could investigate how the sentiment scores were effected by controversial actions of a certain politician. In this example the metric change over time of the politician Maria Vassilakou is analysed.

In figure 8.5 the normalised frequency distribution and sentiment score distribution over time of Maria Vassilakou is shown. The four marked sections outline certain events related to Maria Vassilakou as well as controversial decisions related to her role as Vice-mayor.

1. a) "2004" elected to the federal executive committee of the Green Party
   b) "2005" top candidate of the Greens for the municipal elections
2. "2010" top candidate in the state parliament and municipal council election - elected as vice mayor
3. "2015" state parliament and municipal council election – controversy surrounding her declaration to resign if the Green Party loses vote shares
4. a) "2017" controversial high-rise project at the Heumarkt in Vienna; UNESCO sets the City of Vienna onto the "Red List of World Heritage in Danger"
   b) "2018" announcement that she will not run in the next state parliament and municipal council election

The figure 8.5 shows that during the above described actions of the politician a change in the sentiment scores of different time frames happened. It is important to mention that correlation does not imply causation.

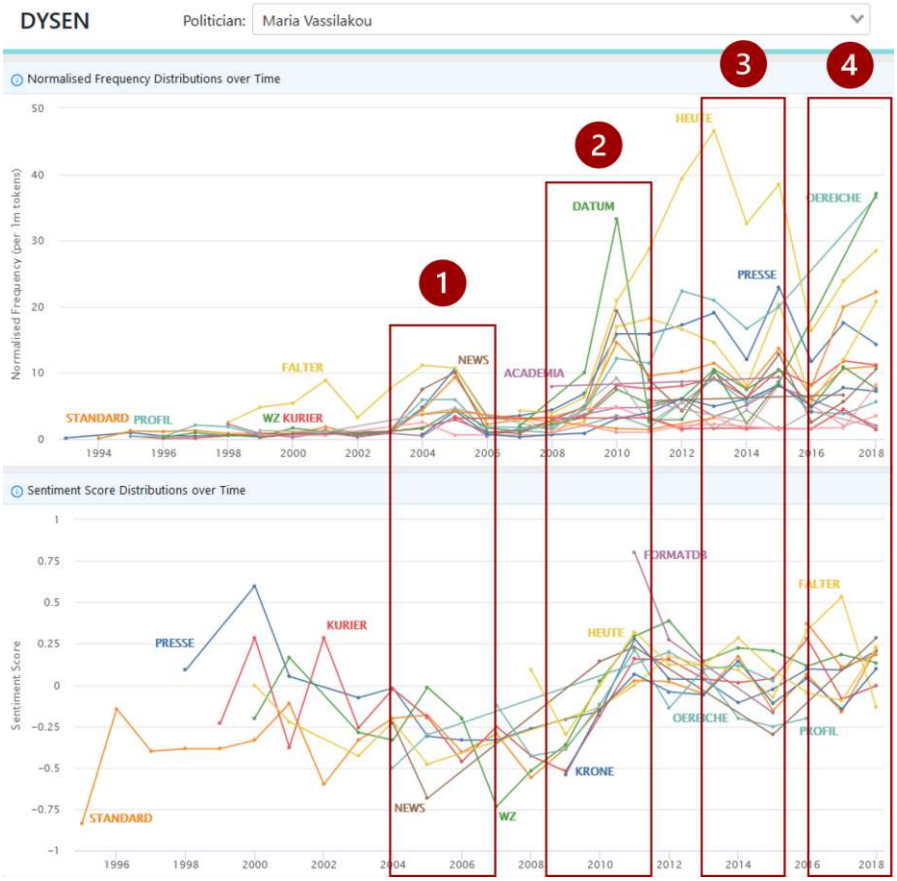

Figure 8.5: A qualitative view on the normalised frequency distribution and sentiment scores over time of Maria Vassilakou by using the web application (Link:https://dysen-tool.acdh-dev.oeaw.ac.at/)

8.5 Ethical Questions

Another important issue is the possibility of abusage of such a system whereby this question is not limited on this work. It is a more general question in the direction: can we trust a system that makes predictions based on a classifier? It is important to always check how a system or model was built and what was e.g. the definition of "bias", what data-sets were used to train the classifier and whether there were some other constraints, etc. In this work the entire process from the beginning to the final classifier is described in detail. This allows a critical look at what constraints are present and what assumptions were met.

CHAPTER 9

Conclusion

9.1 Summary

Today's world is full of stories about scandals such as the BVT affair[1], the Ibiza affair[2] and many other controversial events involving public interest figures, often politicians. Do news media report in a neutral way about persons of interest, especially politicians? Is there a certain bias in news coverage against politicians? This thesis aims to answer that questions by developing a method to investigate that type of bias. In addition this work shows to what extent this bias is detectable. A literature review was performed to retrieve the stat-of-the-art in this narrow domain of sentiment analysis in news media. The literature review needed to cover different fields whereby in addition to the methods applied in regard of machine learning and lexical based approaches also research into the direction of bias in general and political polarization was required. Austrian German and the analysis over time and different news media added additional complexity to this approach. To align this thesis with the DYSEN[3] research project into which the thesis is embedded the focus was limited to Viennese politicians whereby non Viennese or non Austrian wide news media were excluded (e.g. Vorarlberger Nachrichten (VN), ...). During the research it became increasingly clear that lexical approaches must also be included in the work. Topics like explainability and transparency are currently heavily researched and very important. Supervised approaches e.g. deep learning methods in general are not transparent and often not explainable.

To answer the questions raised, before the collection of as much as possible data, was required. This was done by using the with linguistic data by Ransmayer et al. [JKM17] enriched Austrian media corpus (AMC)[4] which covers nearly the whole Austrian media

[1] https://de.wikipedia.org/wiki/BVT-Affäre
[2] https://en.wikipedia.org/wiki/Ibiza_affair
[3] https://dylen.acdh.oeaw.ac.at/dysen/
[4] https://amc.acdh.oeaw.ac.at/

landscape. Additional data sources were added to enhance the lexical based approach by using a data-set from STANDARD by Schabus et al. [SST17]. To additionally incorporate the regional aspects of the Austrian German an Austriacism list by Ammon et al. [ABE16] combined with a list based on Wikipedia[5] was added as third data-source.

For both, the supervised and the lexical based approaches, data extraction and data labeling was required. For the data extraction phrases around Viennese politicians[6] were extracted out of the corpora. The STANDARD data-set is already labeled which did not required additional labeling. During the labeling process crowd sourcing was conducted. The crowd sourcing was performed by the project colleagues over Prolific[7] as management platform and SoSci Survey[8] as survey platform. The data preparation and evaluation of the results was performed as part of this thesis. For each of the methods, the lexical and the supervised approaches multiple internal and external crowd sourcing steps were required. Data quality was ensured by implementing golden samples whereby the annotators needed to reach 75% correctness. A majority vote was built and the inter-annotator agreement was evaluated. For the BWS based approach proposed by Kiritchenko & Mohammad [KM17a, KM17b] which required labeling of tuples by selecting the most positive and most negative word a split-half reliability was calculated.

For the lexical approach the current state-of-the-art in regard of sentiment lexicons was evaluated. Due to the lack of German sentiment dictionaries in this domain which was researched by Kern et al. [KBK+21] a new language resource was created which is called "Austrian Language Polarity in Newspapers (ALPIN)". For ALPIN two different sentiment score algorithms were utilized on the one hand SPLM by Almantarneh & Gamallo [AG18] for the AMC & STANDARD data-sets and on the other hand BWS for the crafted Austriacism list. Postprocessing was required for aligning the resulting data-sets of both sentiment score algorithms by scaling them to a common range of [-1,+1]. The evaluation of the crafted language resource was done by evaluating the crafted sentiment dictionary against the labeled data-set and splitting up the labeled data-set into a train and test data-set. The train/test split was performed before applying the SPLM algorithm to prevent dependencies between the language resource and the test data-set. A second evaluation was performed by using two external sentiment dictionaries and checking their effectiveness against the labeled data-set. The used sentiment dictionary were "Affective Norms (AN)" by Köper & Schulte im Walde [KSiW16] and "German Polarity Clues (GPL)" by Waltinger [Wal10]. At the end the word intersections which occurred by merging the AMC and Austriacism list data were analyzed and compared.

The supervised approaches were conducted by using the labeled AMC data-set for training the classifiers. First a baseline by training different common classifiers was performed. The results of this baselines were compared by calculating different metrics like "precision", "recall", "f1 score" and "accuracy". After this baseline more complex approaches were

[5] https://de.wikipedia.org/wiki/Liste_von_Austriazismen
[6] https://www.wien.gv.at/kultur/archiv/politik/
[7] https://prolific.co/
[8] https://www.soscisurvey.de/

applied. FastText by Joulin et al. [JGBM16] and BERT by Devlin et al. [DCLT19] were utilized. FastText retrieved good results by low cost in terms of computational resources. BERT based algorithms are currently part of the state-of-the-art in this domain. Therefore two different pre-trained BERT models were selected and fine-tuned with the labeled data-set of the AMC. Evaluation has been shown that both BERT based approaches outperform all the other methods including the lexical based approach. Not only the common metrics accuracy, precision, recall and f1 score are factors which need to be considered. Also explainability and transparency are important metrics which need to be kept in mind. For the web application[9], created by the project team, the lexical based model is used. In the end the results were evaluated qualitatively by showing possible usages.

9.2 Contribution

In this section the contributions to the state-of-the-art are explained. The first part shows the two research questions that were answered by this work. The second part lists the publications related to this work.

9.2.1 Research Questions

- **RQ1:** To what extent is it possible to predict the polarization of politicians over time in different media?
- **RQ2:** How well do different approaches of machine learning perform in predicting the polarisation of politicians in the context of sentiment analysis in the Austrian news media?

ad. RQ1: This question is addressed qualitatively by the web application and quantitatively by the evaluation of the used models and algorithms. The results confirm that tendencies and dynamics can be captured well.

ad. RQ2: State-of-the-art ML models perform well on this task and show better result as traditional dictionary-based approaches. However, with our created language resource good results are achieved and explainability is enhanced. In addition, it is available publicly[10] (thus can be easily used also by researchers from the humanities and social scientists).

This work makes the different sub-types as described by Eberl et al. [EBW17] accessible. Visibility is represented by the relative frequency of politicians across media and time. Tonality is made available through the calculated sentiment scores towards politicians. Agenda is represented by the imbalance between different political directions. This imbalance can be accessed through the web application.

[9] https://dysen-tool.acdh-dev.oeaw.ac.at/
[10] https://doi.org/10.5281/zenodo.5857150

Several models were created. On the one hand models based on the developed sentiment dictionary which is the first sentiment dictionary for this domain. This dictionary consists of several parts, in particular the AMC based part and the Austriacism list are the first of their kind in this domain. During the data annotation the Best-Worst Scaling approach was combined by performing crowd-sourcing instead of using fixed annotators. This novel approach was very successful and lead to results with a high split-half reliability. On the other hand different machine learning based models were created whereby especially the BERT fine tuned models are enhancing the current state-of-the-art for this domain.

9.2.2 Publications

Kern, B. M., Baumann, A., Kolb, T. E., Sekanina, K., Hofmann, K., Wissik, T., & Neidhardt, J. (2021a). A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis. In 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik. **Shortlisted for Best Paper** https://doi.org/10.4230/OASIcs.LDK.2021.37

Submitted to LREC 2022 (under review):
Kolb, T. E., Kern, B. M., Sekanina, K., Wissik, T., Neidhardt, J., Baumann, A., (2022) The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers.

Data-set:
Kolb, Thomas Elmar, Sekanina, Katharina, Kern, Bettina M. J., Neidhardt, Julia, Wissik, Tanja, & Baumann, Andreas. (2022). The ALPIN Sentiment Dictionary: Austrian Language Polarity in Newspapers (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5857151

9.3 Future Work

Considering the wide range of different areas this thesis deals with, there are limitations in some fields. The main limitation of this work is caused by the limited financial budget and the limited allowed token count during text extraction out of the Austrian media corpus. With a higher budget labeling of a bigger reference data-set would have been possible. This would also open up the possibility of expanding the scope of observed politicians from Vienna to a broader range and labeling a data-set for a more sophisticated method such as aspect-based sentiment analysis [HBR19]. During evaluation the problem arose that there is currently no labeled data-set in this narrow domain that could be used as an external test data-set. Last but not least, depending on the method, there is room for further research in the area of explainability and transparency, which are already addressed in this work but have not yet been done, for the deep learning based algorithms.

List of Figures

1.1 DYSEN web application 3
1.2 workflow 5

5.1 Direct annotation survey: comparison of different “b” settings 34

6.1 ALPIN - workflow 39
6.2 AMC & STANDARD data-set scaling - comparison 47
6.3 Austriacism list scaling - comparison 47

8.1 Politician distribution in the overall news media landscape 65
8.2 Top 15 politicians over all news media 66
8.3 Normalised frequency distributions over time of Werner Faymann (Link: https://dysen-tool.acdh-dev.oeaw.ac.at/) 67
8.4 Sentiment Score Distributions over time of Werner Faymann (Link:https://dysen-tool.acdh-dev.oeaw.ac.at/) 67
8.5 A qualitative view on the normalised frequency distribution and sentiment scores over time of Maria Vassilakou by using the web application (Link:https://dysen-tool.acdh-dev.oeaw.ac.at/) 69

List of Tables

4.1 Unique politician list based on POLAR after filtering. 17
4.2 APA & OTS press releases based on AMC v3.1. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17]. 17
4.3 News media list based on AMC v3.1 part 1/2. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17]. 18
4.4 News media list based on AMC v3.1 part 2/2. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17]. 19
4.5 TV media list based on AMC v3.1. Table data retrieved and column names shortened from https://amc.acdh.oeaw.ac.at/dokumentation/medienliste/ which is based on the work of Ransmayr et al. [JKM17]. Remark: "MWVOLL" is the Austrian public service broadcaster called "Österreichischer Rundfunk (ORF)". 19
4.6 AMC politician phrases with lemma, PoS and negation tagging 23
4.7 Tag to WordNet® word-form assignment to harmonise the tagging 23
4.8 AMC phrases based on Viennese politicians after preprocessing 24
4.9 STANDARD posts data-set structure 25
4.10 Overview of the STANDARD posts data-set 26
4.11 Tag to WordNet® word-form assignment to harmonise the tagging 26
4.12 Standard posts after preprocessing 27
4.13 Overview of the Austriacisms data-set 27

5.1 Kappa statistic categorization proposed by Landis & Koch [LK77] 30
5.2 Direct annotation survey 34
5.3 Main survey golden samples 36
5.4 Overview of the Austriacisms labeling result after performing the main survey 36

6.1 AMC labeled data-set (5.315 rows) 41
6.2 AMC data-set after applying SPLM: positive words (2.368 rows) 42
6.3 AMC data-set after applying SPLM: negative words (2.448 rows) 42
6.4 STANDARD labeled data-set (3.599 rows) 43
6.5 STANDARD data-set after applying SPLM: positive words (2.718 rows) . 43

6.6 STANDARD data-set after applying SPLM: negative words (2.399 rows) . 44
6.7 AMC & STANDARD data-set combined (8.914 rows) 44
6.8 AMC & STANDARD combined after applying SPLM positive words (4.605 rows) 45
6.9 AMC & STANDARD data-set combined after applying SPLM: negative words (4.324 rows) 45
6.10 Austriacism list (537 rows) 46
6.11 Combined data-set with Sentiment scores based on SPLM & BWS 48
6.12 German Polarity Clues (9.561 rows) 49
6.13 Affective Norms (351.502 rows) 50
6.14 ALPIN against STANDARD data-set 50
6.15 ALPIN against AMC data-set 51
6.16 Results STANDARD data-set (train-test) 51
6.17 Results AMC data-set (train-test) 52
6.18 Results STANDARD & AMC data-set (train-test) 52
6.19 Top and bottom 10 word intersections: Austriacisms vs. AMC & STANDARD data-set (32 rows in total) 53

7.1 AMC data-set after preprocessing and filtering out neutral items (2.638 rows in total) 59
7.2 Baseline results of the test data-set (= average over all folds) 60
7.3 AMC data-set after preprocessing and filtering out neutral items (2.638 rows in total) 60
7.4 FastText results split by label 61
7.5 BERT (distilbert-base-german-cased) fine-tuned with the AMC data-set results by step 62
7.6 BERT (dbmdz/bert-base-german-cased) fine-tuned with the AMC data-set results by step 62
7.7 Comparison “dbmdz/bert-base-german-cased” (DBMDZ) and “distilbert-base-german-cased” fine-tuned (DBGC) 62

8.1 Top 15 politicians over all news media 66

Bibliography

[ABE16] Ulrich Ammon, Hans Bickel, and Jakob Ebner. *Variantenwörterbuch des Deutschen : die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen.* Walter de Gruyter, Berlin, 2016.

[Ada10] Sean Aday. Chasing the Bad News: An Analysis of 2005 Iraq and Afghanistan War Coverage on NBC and Fox News Channel. *Journal of Communication*, 60(1):144–164, 2010.

[AG18] Sattam Almatarneh and Pablo Gamallo. Automatic Construction of Domain-Specific Sentiment Lexicons for Polarity Classification. pages 175–182, June 2018.

[Amm95] Ulrich Ammon. *Die deutsche Sprache in Deutschland, sterreich und der Schweiz: Das Problem der nationalen Varietäten.* de Gruyter, 1995.

[BC18] Kevin K. Banda and John Cluverius. Elite polarization, party extremity, and affective polarization. *Electoral Studies*, 56:90–101, 2018.

[BKBH21] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, August 2021.

[BKKK16] Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. Opinion mining and sentiment analysis. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 452–455, 2016.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[BNH17] Penubaka Balaji, O. Nagaraju, and D. Haritha. Levels of sentiment analysis and its challenges: A literature review. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, pages 436–439, March 2017.

[BS16] Gerhard Backfried and Gayane Shalunts. Sentiment Analysis of Media in German on the Refugee Crisis in Europe. In Paloma Díaz, Narjès Bellamine Ben Saoud, Julie Dugdale, and Chihab Hanachi, editors, *Information Systems for Crisis Response and Management in Mediterranean Countries*, pages 234–241, Cham, 2016. Springer International Publishing.

[CDK+06] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online Passive-Aggressive Algorithms. *J. Mach. Learn. Res.*, 7:551–585, December 2006. Publisher: JMLR.org.

[CL11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), May 2011. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[CPE22] Rosario Catelli, Serena Pelosi, and Massimo Esposito. Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. *Electronics*, 11(3), 2022.

[DBLJ14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. July 2014.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. _eprint: 1810.04805.

[DDKO14] Michael Dimock, Carroll Doherty, Jocelyn Kiley, and Russ Oates. Political polarization in the American public. *Pew Research Center*, 12, 2014.

[DMGDlP20] Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*, 9(3), 2020.

[DPMA19] Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications*, 118:272–299, 2019.

[EBW17] Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences. *Communication Research*, 44(8):1125–1148, 2017. Publisher: SAGE Publications Inc.

[FA08] Morris P. Fiorina and Samuel J. Abrams. Political Polarization in the American Public. *Annual Review of Political Science*, 11(1):563–588, June 2008. Publisher: Annual Reviews.

[Fle74] Joseph L. Fleiss. Statistical methods for rates and proportions. 1974.

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, May 2010. PMLR.

[HBR19] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-Based Sentiment Analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland, September 2019. Linköping University Electronic Press.

[HC10] Edward S Herman and Noam Chomsky. *Manufacturing consent: The political economy of the mass media*. Random House, 2010.

[Hev07] Alan Hevner. A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19, January 2007.

[HJ17] Martin Haselmayer and Marcelo Jenny. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & Quantity*, 51(6):2623–2646, November 2017.

[HMVLB20] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. original-date: 2014-07-03T15:15:40Z.

[HWM17] Martin Haselmayer, Markus Wagner, and Thomas M. Meyer. Partisan Bias in Message Selection: Media Gatekeeping of Party Press Releases. *Political Communication*, 34(3):367–384, July 2017. Publisher: Routledge.

[JGBM16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*, 2016.

[JKM17] Jutta Ransmayr, Karlheinz Mörth, and Matej Ďurčo. *II. AMC (Austrian Media Corpus) - Korpusbasierte Forschungen zum Österreichischen Deutsch*. Digitale Methoden der Korpusforschung in Österreich. Verlag der Österreichischen Akademie der Wissenschaften, Wien, 2017. 27.

[KBK+21] Bettina M. J. Kern, Andreas Baumann, Thomas E. Kolb, Katharina Sekanina, Klaus Hofmann, Tanja Wissik, and Julia Neidhardt. A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis. In Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch, editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*,

volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 37:1–37:17, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 2190-6807.

[KJ20] Akshi Kumar and Arunima Jaiswal. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, 32(1):e5107, January 2020. Publisher: John Wiley & Sons, Ltd.

[KM17a] Svetlana Kiritchenko and Saif Mohammad. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[KM17b] Svetlana Kiritchenko and Saif M. Mohammad. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best-Worst Scaling, 2017. _eprint: 1712.01741.

[KSiW16] Maximilian Köper and Sabine Schulte im Walde. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[Lan11] Pat Langley. The changing science of machine learning. *Machine Learning*, 82(3):275–279, 2011.

[Lel16] Yphtach Lelkes. Mass Polarization: Manifestations and Measurements. *Public Opinion Quarterly*, 80(S1):392–410, January 2016.

[Liu15] Bing Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.

[LK77] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. Publisher: [Wiley, International Biometric Society].

[McD09] Gary C. McDonald. Ridge regression. *WIREs Computational Statistics*, 1(1):93–100, 2009. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wics.14.

[MGK16] Mika Mäntylä, Daniel Graziotin, and Miikka Kuutila. The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. *Computer Science Review*, 27, 2016.

[MHK14] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[NWCG10] Thomas R Nichols, Paola M Wisner, Gary Cripe, and Lakshmi Gulabchand. Putting the Kappa Statistic to Use. *The Quality Assurance Journal*, 13(3-4):57–61, July 2010. Publisher: John Wiley & Sons, Ltd.

[NY03] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, Sanibel Island, FL, USA, 2003. Association for Computing Machinery. Type: 10.1145/945645.945658.

[Orm09] Bryan K. Orme. MaxDiff Analysis : Simple Counting , Individual-Level Logit , and HB. 2009.

[Per21] Denilson Alves Pereira. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115, February 2021.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Rai13] Prashant Raina. Sentiment Analysis in News Articles Using Sentic Computing. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 959–962, 2013.

[RHW+18] Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, 12(2-3):140–157, April 2018. Publisher: Routledge.

[RTBE18] Jacobo Rouces, Nina Tahmasebi, L. Borin, and Stian Rødven Eide. Generating a Gold Standard for a Swedish Sentiment Lexicon. In *LREC*, 2018.

[SD21] Shashank Shekhar Sharma and Gautam Dutta. SentiDraw: Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination. *Inf. Process. Manag.*, 58:102412, 2021.

[SDCW20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, February 2020. arXiv: 1910.01108.

[Sid19] Wladimir Sidorenko. Sentiment Analysis of German Twitter. *CoRR*, abs/1911.13062, 2019. arXiv: 1911.13062.

[SST17] Dietmar Schabus, Marcin Skowron, and Martin Trapp. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1241–1244, New York, NY, USA, 2017. Association for Computing Machinery. event-place: Shinjuku, Tokyo, Japan.

[SVA19] Wataru Souma, Irena Vodenska, and Hideaki Aoyama. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46, January 2019.

[tea20] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[TL16] Charles S. Taber and Milton Lodge. The Illusion of Choice in Democratic Politics: The Unconscious Impact of Motivated Political Reasoning. *Political Psychology*, 37(S1):61–85, 2016.

[vAvdVB21] Wouter van Atteveldt, Mariken A. C. G. van der Velden, and Mark Boukes. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2):121–140, April 2021. Publisher: Routledge.

[Wal10] Ulli Waltinger. GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010. electronic proceedings.

[WDS+20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma,

Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[XGR20] Ning Xie and Derek Doran Gabrielle Ras, Marcel van Gerven. Explainable Deep Learning: A Field Guide for the Uninitiated. *arXiv*, 2020.

[YV20] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, August 2020.

[Zho17] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, August 2017. _eprint: https://academic.oup.com/nsr/article-pdf/5/1/44/31567770/nwx106.pdf.