Informatics

Explainability in Hate Speech Detection

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Markus Reichel, BSc
Matrikelnummer 01529191

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Allan Hanbury
Mitwirkung: Gábor Recski, PhD
Kinga Gémes, BSc MSc

Wien, 22. Juli 2022

Markus Reichel

Allan Hanbury

Informatics

Explainability in Hate Speech Detection

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Markus Reichel, BSc
Registration Number 01529191

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dr. Allan Hanbury
Assistance: Gábor Recski, PhD
Kinga Gémes, BSc MSc

Vienna, 22nd July, 2022

Markus Reichel   Allan Hanbury

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

Erklärung zur Verfassung der Arbeit

Markus Reichel, 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. Juli 2022

Markus Reichel

Danksagung

Ich möchte hiermit Gabriele, Helga und Ernst Reichel danken, ohne deren Unterstützung ich kaum in der selben Zeit meine beiden Studien vorantreiben hätte können.

Außerdem danke ich meinem Betreuer Gábor Recski für die Möglichkeit, direkt an der Forschung an den Themen Erklärbarkeit, Deep Learning und Natürlicher Sprachverarbeitung zu arbeiten, Kinga Gémes, Ádám Kovács und Martin Bär für die tolle Zusammenarbeit und Philipp Adam und Samantha Fuchs fürs Korrekturlesen.

Acknowledgements

I would like to thank Gabriele, Helga and Ernst Reichel, without whose support I would have hardly been able to advance in my two studies in the same time.

Additionally, I thank my advisor Gábor Recski for the opportunity to directly work in the research fields of explainability, deep learning and natural language processing, Kinga Gémes, Ádám Kovács and Martin Bär for the great cooperation and Philipp Adam and Samantha Fuchs for proof-reading.

Kurzfassung

Diese Arbeit beschäftigt sich mit Methoden der Erklärbarkeit von Natürlicher Sprachverarbeitung am Beispiel vom Problem der Erkennung von Hassreden und beleidigenden Inhalten. Nachdem eine Einführung in das Hassrede-Problem gegeben wurde, wird zuerst ein Paper über unsere Baseline-Systeme für solch eine Gemeinschaftsaufgabe studiert. Danach werden die Konzepte der sogenannten Rationale und auf Rationalen basierenden Erklärbarkeitsmetriken präsentiert, welche anschließend benutzt werden, um nicht nur die Performance, sondern auch die Erklärbarkeitsmetriken Plausibilität und Treuhaftigkeit von drei Deep-Learning Modellen mit denen von händisch erstellten regelbasierten Systemen auf den beiden Aufgaben der Erkennung von Texten, die Frauen und Homosexuelle angreifen, zu vergleichen. Für diese Aufgaben wird der Datensatz HateXplain in kleinere Datensätze aufgeteilt, die für die Erkennung von Hass und Beleidigungen gegen jene Zielgruppen erzeugt wurden. Zudem werden auch die menschlichen Annotationen bezüglich ihrer Erklärbarkeit zum Vergleich ausgewertet. Am Ende wird eine qualitative Fehleranalyse durchgeführt.

Wir lernen, dass Regeln besser in der Präzision und Treuhaftigkeit performen und Deeplearning-Modelle im F1-Score, manche menschlich-annotierten Rationale nicht unbedingt als Gold-Label betrachtet werden sollten und gut-performende Regeln nicht notwendigerweise auch Regeln sind, welche gut-erklärende Rationale zurückgeben. Dennoch kann man sagen, wenn jene Regeln in solch einer Form entwickeln wurden, dass sie gute Rationale berechnen, dann kann deren Leistung der Erklärbarkeit höher sein als jene von Deep-Learning Modellen, mit und ohne Attention-Mechanismus.

Warnung: Diese Arbeit enthält beleidigende Wörter.

Abstract

This work examines the explainability of natural language processing on the example of hate speech and offensive content detection. After an introduction to the hate speech task is given, first a paper about our baseline systems on such a shared task is reviewed. Afterwards, the concepts of rationales and rational-based explainability metrics are presented, which are then used to compare not only the performance but also the explainability-metrics plausibility and faithfulness of three deep learning models with those of hand-made rule-based systems on the two tasks of detecting offensive text targeted against women and homosexuals. For these tasks, the dataset HateXplain is processed into smaller datasets specifically for detecting hate and offensive content against these specific target groups. Also, human annotations are evaluated in terms of their explainability for comparison. In the end, an qualitative error analysis is conducted.

We learn that rules perform better in precision and faithfulness and deep learning models in F1 score, some human-annotated rationales should not necessarily be viewed as gold labels and well-performing rules are not necessary rules which yield well explaining rationales. However, if the rules are engineered in a way to predict good rationales, explainability performance can be higher than deep learning models with and without attention-based mechanisms.

Disclaimer: This work contains profane words.

Contents

CHAPTER 1

Introduction

We start by presenting the problem statement and the expected results of this thesis, where we give information about the hate speech detection task and the dataset we are using. Then, the target audience and the contribution of others is described. Finally, we show how this work is structured.

1.1 Problem statement

Identifying hate speech and offensive content is a highly relevant Natural Language Processing (NLP) problem of online platforms. Regardless if it is a forum, social network or other form of online community, if platform activity reaches a certain amount, moderating its content by manual means becomes unfeasible. Therefore, several detection models are developed to assist in the moderation of unwanted posts, comments and alike. Offensive content detection is typically defined as a classification task in most datasets like HASOC [has19], hatEval [hat19] and Germeval [ger19]. Classification can be a binary distinction between offensive and non-offensive content, but also adding more refined classes makes sense, e.g. one for clear personal offence, one for swearing and bad language etc. Table 1.1 shows example data of the problem. On the other hand, lately, a lot of deep learning models became state of the art (see Section 2). Deep learning models do not rely on manual feature engineering, as the model is able to extract the features itself. However, the drawback of these models come with the reduced explainability, as they are black box models. But it makes not only from a technical and ethical perspective sense to get an explanation for a decision made by the model. It is also enforced by law in several countries to give citizen the right to get an explanation for certain decisions, like in the European Union (see Goodman and Flaxman [GF17]).

Classes	Text
NOT	#ShameOnICC #iccworldcup2019 world cup reality https://t.co/MhPD5gDVze
HOF,PRFN	This is everything. #fucktrump https://t.co/e2C48U3ps
HOF,HATE	@TheRealOJ32 You belong in jail and then in hell #murderer #JusticeForNicoleAndRon
HOF,OFFN	#Trump just confessed to being a traitor - #TrumpIsATraitor https://t.co/EkGv0T02fN

Table 1.1: Text examples with their different class tags from Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages 2020 (HASOC 2020) [has20]. In this shared task, the high-level distinction was between hate-and-offensive (HOF) and not hate speech (NOT), while the HOF class was further divided into profane (PRFN), hate speech (HATE) and offensive (OFFN).

1.2 Expected results

First of all, a baseline model should be created on an appropriate hate speech dataset. This means that an existing easy-to-setup system, typically a standard architecture, which performs comparable to current state-of-the-art models for hate speech tasks, should be trained on the data to get a model to work with and do further experiments on. A prominent classic architecture is e.g. the Support Vector Machine (SVM) by Suthaharan and Shan [Sut16], a prominent deep learning architecture would be Bidirectional Encoder Representations from Transformers (BERT) by Devlin et al. [DCLT18]. After this baseline model has been developed, an error analysis should be conducted to identify possible problems and points of improvement of the setup. Existing tools from the research group will be used to extract rules, mainly the exPlainable infOrmation exTrAcTion framewOrk (POTATO) from Ádám Kovács et al. [KGIR22]. The rules are generated by hand and/or statistically analyzing which text patterns are most common among the target classes. They will be used for the error analysis and as a starting point to get more insight into the problem and how a globally explainable, high-precision rule system can be created.

1.2.1 Research questions

This brings us to the following research questions:

1. How well does the baseline model trained on the hate speech dataset perform in terms of precision and recall?
2. What characterizes typical false positives and false negatives, and what linguistic patterns characterize the errors?
3. What syntactic and semantic rules can be formulated to improve the rule-based model?
4. How well do rules developed for the dataset perform compared to the machine-learning baseline?

1.2.2 Datasets

Initially when the thesis proposal was created, the Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) challenge was chosen as a promising source for training and evaluation data [has21]. From 2019 onward, every year datasets in several languages are published [has20] [has19]. The challenge is divided into two subtasks, the first one being a binary distinction between hate-and-offensive and non-hate-and-offensive, and the second one with a finer distinction of the hate-and-offensive class in profane, offensive and hate speech. The initial research questions were also formulated on the HASOC dataset. The participation in the HASOC 2021 challenge led to the development of a preliminary paper capable of answering the research questions for the HASOC dataset, which is done in the end of Chapter 2. There, we can already see that rule-based systems perform better in precision, while deep-learning systems perform better in F1-score.

However, after our participation in the HASOC 2021 challenge, the HateXplain dataset by Mathew et al. [MSY^{+}21] was discovered by our research group. This new English dataset had various new concepts like different target classes of the offensive posts ("Women", "Homosexual", "Black", etc.) and especially rationale annotations. With this rationales, it is possible to test the explainability of a model. It has therefore been chosen as the main dataset. Experiments on this dataset did not only confirm findings from our HASOC 2021 paper, but also showed that rules have a better faithfulness-explainability score than deep learning models, well-performing rules do not necessarily return well explaining rationales, and some controversial human annotations of the dataset interfere with our evaluation.

1.3 Target audience of the thesis

This thesis is written for two different kind of readers. The first type is an expert in the field of the hate speech detection task, whom we want to present our new results and approaches of comparing deep learning and rule-based models in aspects of performance and explainability. The second type of reader might be adept in some fields of computer science and informatics but does not have a lot of experience in the artificial intelligence (AI) task of designing and/or training models, but wants to deepen his or her knowledge in this field by looking at a practical application capable of showing state-of-the-art methods. We believe that studying the hate speech detection task is a good way of achieving this. Knowledge of the fundamental concepts of machine-learning is assumed, for interpreting the performance metrics of our models, we especially assume that the reader has knowledge about the concepts of precision, recall and F1 score. Further, the concepts of supervised learning for our deep learning models and graphs for our rule-based systems are important.

1.4 Contribution of others

The participation in the HASOC 2021 challenge was done in a team consisting of Kinga Gémes, Ádám Kovács and me, Markus Reichel, with supervision of Gábor Recski. For the first task, Ádám created a rule system semiautomatically with the framework POTATO, while Kinga and I respectively developed a deep learning baseline by fine-tuning the BERT architecture. The two BERT variants performed in a similar range. Due to some differences like casing and an additional balanced loss function, Kinga's BERT variant performed better and was used for the final submission together with Ádám's rules. Kinga also experimented on the second task with both BERT and Random Forest approaches.

For experiments on the HateXplain dataset, another Team consisting of Kinga Gémes, Martin Bär and me was formed, again with supervision by Gábor Recski. This time, two rule systems for two different target classes were created. Kinga Gémes created the rules for target "Women" and some of the data preprocessing required for separating the HateXplain dataset into different sets and target classes, while Martin Bär created the rules for target "Homosexual". Everything else covered in this thesis was done by me, Markus Reichel.

1.5 Structure of this work

This work is organized as follows: After this introduction Chapter 2 starts by presenting related literature and several topics used in this thesis and ends by reviewing a preliminary paper from us which laid the groundwork for developing our first hate speech detection baseline model and answers the research questions for the HASOC dataset. Chapter 3 then continues by describing selected methods for achieving our research goals for our main dataset HateXplain. Experimental setup and results are presented in Section 4. Then, Chapter 5 continues by presenting the qualitative error analysis. Finally, Chapter 6 complements the thesis by giving an outlook to future work and then a summary of this thesis while drawing conclusions.

CHAPTER 2

Related Work

In this chapter we present related work and at the end also preliminary work done by us.

2.1 Datasets for the hate speech detection task

We begin by giving an overview of literature about commonly used hate speech and offensive content datasets and tasks. Often, open accessible datasets are created in the context of a shared task with an open leaderboard, where people and teams can participate in building the best performing system. Therefore, the task overview papers following the task submission deadline represent a direct source for state-of-the-art systems.

2.1.1 Classic hate speech and offensive content datasets and tasks

Recent offensive content detection tasks include HASOC, hatEval and Germeval. They were chosen to study the current state-of-the-art. It has to be noted that the actual definition of hate speech and the according labels for classifcation can be different between datasets. Further information on the labels specifically for HASOC will be presented in Section 2.5 where we discuss our submission to HASOC 2021.

Regarding state-of-the-art systems, HatEval 2019's (Basile et al. [BBF$^+$19]) best model used a Support Vector Machine (SVM) with sentence embeddings from Google's Universal Sentence Encoder, while places 3-5 used CNNs & LSTMs. In Germeval 2019, the model which performed best on all three subtasks was BERT based, while a knowledge based system from TU Vienna did very well on the first two tasks (Struß et al. [SSR$^+$19]). According to Mandl et al. [MMKMC20], most of the successful contributions to HASOC 2020 were BERT-based models, like DistilBERT, RoBERTa and ALBERT. The best and second performing German model of task A were using these fine-tuned models, as well as position four. Task B results were similar. However, there were also teams getting great results with other architectures, as the best model for English task A was using

LSTMs and GloVe embeddings. Second place was an ensemble of BERT, LSTM and CNN with majority vote. The best Hindi task A submission used CNNs with Facebook's fastText library. Still, BERT was also very dominant within the scope of those languages, as English and Hindi task B winners were BERT-based too.

A related task to hate speech detection is emotion detection. Emo2Vec from Xu et al. [XMW+18] is a relevant paper which shows that word embeddings, concatenated with normal GloVe vectors, already achieve state-of-the-art results of emotion detection with a simple logistic regression classifier. It also shows how important representation can be. Word encoding techniques like word2vec by Church and Ward [Chu17] are based on neural networks to generate vectors for traditional classification algorithms. These methods help a lot to represent the semantic features, however, they are context independent, which could be a major problem when analyzing whole sentences. Because of this problem, sentence embeddings like SentenceBERT by Reimers et al. [RG19] exist.

2.1.2 HateXplain and other rationale datasets

HateXplain from Mathew at al. [MSY+21] is a dataset of 20,148 posts collected from the platforms Twitter and Gab. It is crowd-annotated with Amazon MTurk and contains 3 classes regarding offensive and hateful content (hate, offensive, normal). It has also a second, so called "target" annotation. There are several target communities like black people, women and LGTBQ. What makes this dataset special in comparison to other offensive data like HASOC is the fact that it contains so called "rationales". These rationales are annotated additionally by the MTurk workers to highlight parts of the sentence as a reason of their choice of classification. Each post contains in total 1 label and at least 1 target from 3 annotators, and rationales from 1-2, at most 3 different annotators.

HateXplain also tested their dataset on five different deep learning architectures in dimensions of performance, explainability and bias, using the 3-class label and the rationales but not the target annotation. Rationales are predicted either by attention mechanisms or with the Local Interpretable Model-Agnostic Explanations (LIME) algorithm by Ribeiro et al. [RSG16]. While the classification performance was measured by standard metrics of precision and recall, they used the ERASER framework from DeYoung et al. [DJR+19] to evaluate the explainability quality of the rationale predictions. The author's main findings were the fact that models which perform good in classification do not necessarily perform good in terms of the explainability metrics. They make this observation based on their quantitative results, but do not conduct qualitative analysis to explore the reasons.

Regarding other datasets, there are several non-hate speech text datasets containing rationales available. DeYoung et al. link to nine different rationale datasets on their website [DJR+] and make them available for download in the ERASER format, some examples being BoolQ by Clark et al. [Cla19] (question answering for yes/no questions), e-SNLI by Camburu et al. [CRLB18] (natural language explanations for vision-language understanding) and CoS-E by Rajani et al. [RMXS19] (human commonsense explanations for the Commonsense Question Answering (CQA) dataset). Still, we were not able to find other public hate speech datasets containing rationales except HateXplain. However, plenty of traditional hate speech detection datasets existed before, as shown in the last Section.

2.2 Deep learning models

Looking at the shared tasks presented in Section 2.1.1 showed that the newer leaderboards were all dominated by deep learning models. Variants of BERT from Devlin et al. [DCLT18] are popular, easy to train (due to already being pre-trained, they just have to be "finetuned" on a smaller set of data), and perform well on several NLP tasks. They are based on the deep learning transformer architecture from Vaswani et al. [VSP⁺17], which incorporates the so called "attention mechanism" without any recurrence or convolutions Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) use. A lot of new systems are based on the transformer architecture, according to Lin et al. [LWLQ21].

Indeed, while studying NLP model architectures in general, Zhong et al. [ZWM19] state there is a clear trend that researchers are moving from traditional machine-learning to deep learning methods. They also introduced a transformer variant themselves. Some other NLP approaches still rely heavily on RNNs and CNNs. Majumder et al. [MPH⁺19] use RNNs to detect emotion in dialogues, called DialogueRNN. Zhong et al. introduce the knowledge enriched transformer (KET) for emotion detection and compare it to contextual long short-term memory (cLSTM), CNN, CNN+cLSTM, BERT_BASE and DialogueRNN models. On four out of five test sets, their transformer model beats all the others. Note that the base model of BERT performs very close to the specifically trained DialogueRNN from Majumder et al. and beats it on two sets, while being very close to the performance on two other sets.

With respect to the hate speech task in particular, there is a BERT variant specific for detecting hate speech from Awal et al. [ACLM21] called "AngryBERT". However, the good performance of deep learning models is no free lunch. The explainability mechanisms of these models are limited (see the following Section 2.3) due to the black box property (discussed by Castelvecchi and Davide [Cas16]) of deep learning models. It is possible to predict rationales by those systems, not only shown by the models from HateXplain [MSY⁺21], but also by work from Arous et al. [ADY⁺21], who designed the Bayesian network MARTA, which incorporates human rationales as a means of local explanation into attention-based models to outperform various baselines in classification accuracy and explainability. The former one being measured in classic precision, recall, accuracy and F1-score metrics, the latter one being measured by the overlap between predicted and annotated rationales.

2.3 Explainability

Now we want to look further into classifying and measuring explainability. For the former, the framework by Danilevsky et al. will be presented, while for the latter, we will discuss the *plausibility* and *faithfulness* metrics of ERASER in detail.

2.3.1 Explainability framework used in this thesis

Danilevsky et al. [DQA$^+$20] define a framework[1] for categorizing explainability in NLP models. They assign several NLP systems into four different groups: Local Post-Hoc, Local Self-Explaining, Global Post-Hoc and Global Self-Explaining. Table 2.1 shows an overview of these groups.

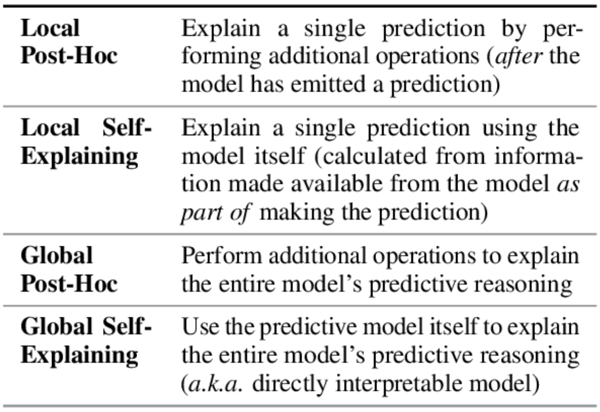
Local Post-Hoc systems explain a single prediction by performing additional operations after the model has made a prediction, like the LIME algorithm. **Local Self-Explaining** models are able to explain a single prediction by the model itself, an example would be feature saliency approaches, like attention mechanisms of deep learning models. **Global Post-Hoc** models perform additional operations to explain the whole model's prediction ability, e.g. SHAP by Lundberg and Lee [LL17], and **Global Self-Explaining** systems use the model itself to explain the whole predictive reasoning. Classic machine-learning models like decision trees, random forests and SVMs would be classified as globally self-explaining.

It is important to note that methods like SHAP can be used for both local and global post-hoc explanation. We can observe that deep learning architectures do not support the global self-explaining class. The self-explainability of attention mechanisms, even if just locally, should also be questioned, as several studies on the validity of this technique found out that the highest attention weights often fail to explain feature importance, as discovered by Serrano and Smith [SS19]. A rationale (which we presented in Section 2.1.2) generated by a deep-learning model is also a local explanation. Rule systems, on the other hand, are globally self-explainable.

2.3.2 Explainability evaluation and ERASER

We already know that the idea behind rationales is to deliver additional local explanation. They are defined as a subset of words from the respective sentence to point out parts of the sentence which might be of greater importance of classifying the input, therefore showing which words are important for the decision. This can be used as an additional means of explainability. ERASER [DJR$^+$19] is a framework for quantitative measurement of the quality of predicted rationales, allowing the evaluation rationale prediction. They define the metrics *plausibility* and *faithfulness* to measure both relevance of the predicted rationales as well as truthfulness in the rationale's influence in the system's decision. It is also used by HateXplain to evaluate the explainability of their baselines.

[1] They also provide an interactive website for visualizing their classification together with the respective papers of their survey at `https://xainlp2020.github.io/xainlp/definitions`.



Local Post-Hoc	Explain a single prediction by performing additional operations (*after* the model has emitted a prediction)
Local Self-Explaining	Explain a single prediction using the model itself (calculated from information made available from the model *as part of* making the prediction)
Global Post-Hoc	Perform additional operations to explain the entire model's predictive reasoning
Global Self-Explaining	Use the predictive model itself to explain the entire model's predictive reasoning (*a.k.a.* directly interpretable model)

Table 2.1: Overview of the high-level categories of explanations [DQA$^+$20, p.3].

But one does not need rationales to measure the explainability of models. Zhang et al. [ZSW22] propose a general way to test the explainability of neural networks by constructing decision trees with a height limit K and test the fidelity of the networks against those. The idea here is the fact that decision trees are both expressive and human-understandable, especially with low height. Due to the decision trees' ability to be also used in NLP (according to Boros et al. [BDP17]), this measure might also be deployed to describe the explainability of NLP classifiers.

One might also question if rationales should be considered as the "gold standard" in the ground truth. Carton et al. [CRT20] test the ERASER faithfulness scores of human-produced rationales on six rationale datasets. They reach the conclusion that those do not necessarily perform well in sufficiency and comprehensiveness, which means that either the annotations fail to capture relevant information, or there are inconsistencies between the human and model task understanding. They also define a way to normalize the two ERASER faithfulness metrics comprehensiveness and sufficiency (for the base definition see Section 3.1.2). Further, they propose two new methods to assess rationale quality, one being based on model training and one using "fidelity curves" to detect properties like redundancy and irrelevance. Jacovi and Goldberg [JG20] reflect on the current state of the faithfulness metric, and come to the conclusion that a binary faithfulness distinction is unproductive due to the nearly non-satisfiability of this assumption as well as the easiness of using a counter-example to disprove an interpretation method's faithfulness. Another conclusion is that human judgement on faithfulness is paradoxical, because the interpretation would be not necessary if they understand the model. Additionally, they state that faithfulness should be evaluated by its own and not together with plausibility, as many authors do not make the distinction between the two, and finally propose guidelines on how evaluation should and should not be conducted.

2.4 Rule-based systems

We have seen that there are several problems with the explainability of deep learning models. As mentioned, they are only locally self-explaining or with additional algorithms like LIME and SHAP locally post-hoc explainable. Therefore, the idea to study globally self-explaining models like rule-based systems in the NLP domain seems promising. Waltl et a. [WBM18] highlight the transparency, readability, maintainability and possibility of directly applying domain knowledge to the rules. Manual labour put into the rules stands for quality. However, requiring domain knowledge and manual labour to create systems can also be seen as a disadvantage, although the biggest limiting factor in their view is the fact that the declarative nature of rules do not generalize that well. Interoperability of rule syntax can be another problem. They also state that machine-learning and rule-based approaches can complement each other well.

POTATO by Kovács et al. [KGIR22] is a task- and language-independent human-in-the-loop (HITL) framework which incooporates a graph-matching rule-based system to classify sentences. It supports the semantic graphs Abstract Meaning Representation (AMR) by Banarescu et al. [BBC+13], Universal Dependencies (UD) by De Marneffe et al. [DMMNZ21] and 4lang by Kornai et al. [KAM+15] for text parsing. HITL means that the user can interactively build and refine rules in the POTATO graphical user interface, while looking at the performance evaluation in real time. POTATO also incooperates machine-learning and therefore a hybrid approach to suggest good rules to the user, and can also be used programmatically as a rule engine. The possible applications span between different domains, examples being German legal text and English social media data.

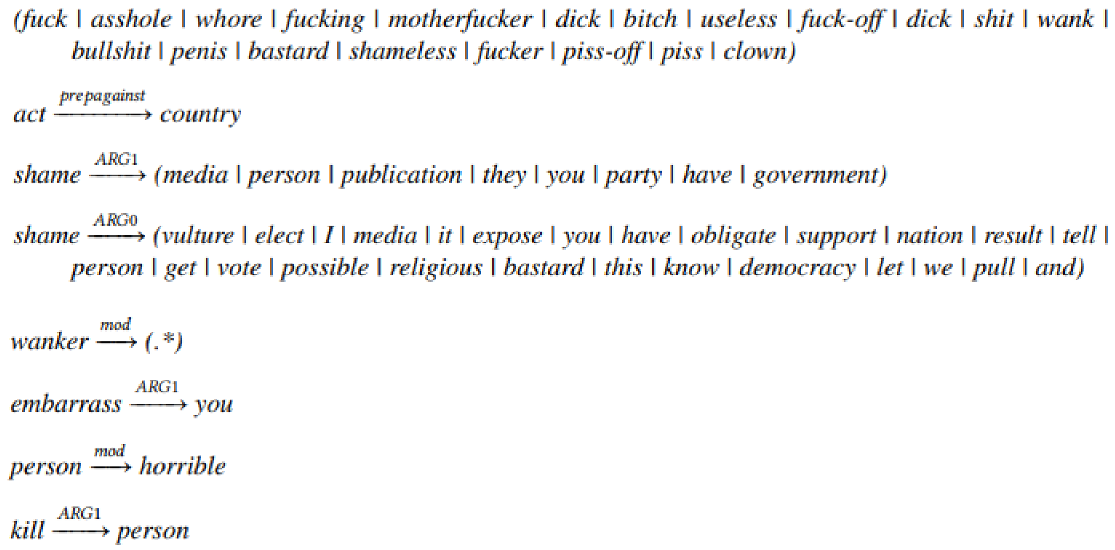
2.5 Review of our submission to HASOC 2021

We now want to summarize our findings during and after our submission to HASOC 2021 [GKRR21]. The paper describes the qualitative and quantitative results of our submitted systems.

2.5.1 Introduction and method

Like their predecessors, HASOC 2021 consists of different subtasks, namely a main task (1a) to distinguish between hate-or-offensive (HOF) and non hate-offensive (NOT) and a fine-grained subtask (1b) between the classes hate speech (HATE), offensive (OFFN) and profane (PRFN) [MMS+21], therefore having the exact same problem statement like HASOC 2020, which was shown in Table 1.1. Our submitted systems were developed for the English language, but the dataset also contains Hindi and (for task 1a) Marathi posts.

At this point, it was already known that leaderboards were dominated by models based on the Transformer architecture [VSP+17], with its most prominent variant BERT [DCLT18]. Due to this architectures having millions of parameters and being hard to interpret, a



(fuck | asshole | whore | fucking | motherfucker | dick | bitch | useless | fuck-off | dick | shit | wank | bullshit | penis | bastard | shameless | fucker | piss-off | piss | clown)

act $\xrightarrow{prepagainst}$ *country*

shame $\xrightarrow{ARG1}$ *(media | person | publication | they | you | party | have | government)*

shame $\xrightarrow{ARG0}$ *(vulture | elect | I | media | it | expose | you | have | obligate | support | nation | result | tell | person | get | vote | possible | religious | bastard | this | know | democracy | let | we | pull | and)*

wanker $\xrightarrow{mod}$ *(.*)*

embarrass $\xrightarrow{ARG1}$ *you*

person $\xrightarrow{mod}$ *horrible*

kill $\xrightarrow{ARG1}$ *person*

Figure 2.1: The complete rule set used in the submission of HASOC 2021. Words represent nodes of the rule subgraph, while arrows and their description represent edges. Note that nodes matching on words may contain regular expressions like (.*) and (|) [GKRR21].

lot of research [DQA$^+$20] is done in the field of explainable AI. One such an approach is globally-explainable rules.

Our final models for task (1a) setup therefore consist on the one hand of one uncased BERT with a single linear classification layer which was fine-tuned by datasets of HASOC 2019 [has19], 2020 [has20], and 2021 [has21]. Metaparameters can be read from our paper [GKRR21]. On the other hand, we created a rule-based system with a HITL approach using the framework POTATO and the 2021 data. Text is parsed into graphs using Abstract Meaning Representation [BBC$^+$13] and rules are also formulated using this representation. By either specifying node- or edge-labeled subgraphs, these graph-rules match on the parsed text and therefore represent a globally-explainable system. The complete rule-system for our submission can be written in under half a page (see Figure 2.1). Also two ensembles of these two models, namely the union of HOF labels for these two, and a logreg voting featuring a logistic regression model with the output of both systems, were submitted.

For subtask (1b), we trained three binary BERTs on each subclass, and also experimented with a Random Forest classifier, but because our focus lies on the first subtask (1a) for answering the research questions (see Section 1.2.1) of this thesis, we will not further go into detail here.

	Offensive			Not offensive			Average		
	P	R	F	P	R	F	P	R	F
NLP-CIC (top)	85.11	90.98	87.95	83.17	73.00	78.15	84.14	81.99	83.05
TUW Logreg voting	81.31	93.23	86.87	85.25	64.60	73.50	83.28	78.91	80.18
TUW BERT-based	80.34	95.24	87.16	88.66	61.49	72.62	84.50	78.36	79.89
TUW Union voting	79.81	95.61	87.00	89.23	60.04	71.78	84.52	77.83	79.39
TUW Rule-based	87.17	45.11	59.45	49.54	89.03	63.66	68.35	67.07	61.56

Table 2.2: Precision (P), recall (R) and F1 score (F) of the top performing submission of HASOC 2021 (NLP-CIC), our logreg ensemble (TUW Logreg voting), our standalone BERT (TUW BERT-based), our rule-based system (TUW Rule-based) and the union ensemble of the two (TUW Union voting) [GKRR21].

2.5.2 Quantitative results

The BERT-based method alone was able to achieve a competitive F1 score, being on third place on the leaderboard regarding the HOF class and ninth place when measuring the average F1-score between HOF and NOT classes. The rule-based system however has the highest precision, and the union of both systems achieved the highest recall. Table 2.2 shows the official test results on HASOC 2021 task (1a).

2.5.3 Qualitative results

The output by both the BERT system and the rule-based system was examined on a randomly selected sample of 150 data points from the HASOC 2019 and 2020 datasets for validation. The used BERT system was the one submitted, while the rules for 2019 and 2020 were separately formulated (also published in our paper [GKRR21]). Generally, four types of data points exist; true positive, false positive, true negative and false negative. The goal of our analysis was to observe the nature of errors (false negatives and false positives) by the two distinct systems as well as the quality and properties of the ground truth annotation of the data points. From the 150 posts, 85 had the gold label NOT, while PRFN, OFFN and HATE (all being HOF) were 37, 15 and 13, respectively, and the focus lied on the binary classification problem of NOT/HOF.

113 tweets were classified correctly by both systems (72 true negatives, 41 true positives). The errors can be classified in three types. Due to BERT having 14 false positives (the first of those is the only post the rule system classified wrongly as positive), this is the first type. The second type are false negatives by both systems, namely 10 samples. 14 additional tweets were missed only by the rule-based system, being false negatives. Therefore, this analysis also points out the high-precision low-recall property of the rule systems.

Regarding the annotation, the first error type of the 13+1 false positive predictions cover sensitive topics and/or contain words used in typical offensive tweets. The second error type with the 10 false negative predictions by both systems include offensive content without using offensive words, or must be viewed in connection with the attached URL to a video to be considered offensive. Sometimes, the reason of the label is unclear to us. Regarding the last type of error (14 false negatives only by the rule system), some of those contain very clear offensive text, while the label of other samples may be questioned. Also, 4 of those 14 posts are probably errors in the data, due to either the use of non-English language or wrongly cut-out offensive content.

2.5.4 Conclusion and retrospective

Two different approaches of creating systems for detecting offensive English tweets were presented, namely a supervised deep learning method and a human-in-the-loop graph feature approach. The BERT model's strength is performing on F1 score and the rule system's is on precision. Also, a detailed error analysis was conducted on a sample of 150 tweets which showed the differences of errors and questioned some of the ground truth of the data.

Working on the HASOC challenge and the paper gave a deep learning baseline and an explainable rule system for the hate speech detection task. This knowledge will later be applied on the HateXplain dataset (starting with Section 3) in this thesis. Additionally, the research questions from Section 1.2.1 can be answered for the HASOC dataset with the qualitative and quantitative results from this preliminary work. While Table 2.2 answers research questions (RQ) 1 and 4, RQ 2 and 3 can be answered by the qualitative results presented in Section 2.5.3 and by Figure 2.1.

CHAPTER 3

Methods

The idea of this work now is to combine the globally self-explaining properties of a rule-based system with the supporting locally self-explaining rationales. In order to accomplish this, the rule based system is extended by the ability to predict such rationales, as well as additional metric calculations for evaluating the explainability of rationale predictions. We then want to compare a deep learning baseline with rule-based systems on the HateXplain dataset in a quantitative and qualitative way. For not only comparing the models in terms of performance but also explainability, the ERASER framework comes into play. This chapter presents our used methods.

3.1 Measuring explainability

Looking further into what exactly the ERASER metrics measure and how they accomplish that is needed for the application on the rule system. As mentioned above, ERASER measures two different aspects of the rationale prediction. Plausibility includes two hard and one soft prediction scores, while faithfulness can be measured by the two scores comprehensiveness and sufficiency.

3.1.1 Plausibility

Plausibility, the first dimension, tries to measure the agreement with human rationales. It can also be interpreted as the following: "How convincing is the interpretation to humans?". ERASER defines two different variants of plausibility: a discrete and a soft selection. For discrete plausibility, two scores are denoted. Intersection-Over-Union(IOU) F1 and Token F1. The idea behind the IOU F1 is to divide the overlap of two different rationale sets by their union. If the result is bigger than a given threshold (e.g. 0.5), it counts as a partial match. The IOU F1 is now the F1-Score of all these partial matches. The continuous version of plausibility aims to measure the metric for continuous rationale predictions. The calculation is done via area under the precision-recall curve (AUPRC). It sweeps also a specific threshold over the token scores, which are continuous in this case.

In order to apply the ERASER framework to rule-based systems, currently, only the discrete plausibility scores are used.

3.1.2 Faithfulness

Faithfulness, on the other hand, aims to measure the influence of the rationales on the prediction. It can be interpreted by the question "How accurately does it reflect the true reasoning process of the model?". Two different faithfulness scores are defined by ERASER, comprehensiveness and sufficiency. To understand these two, we have to define the following expressions.

We define $m(x_i)$ as the probability that the sentence x_i is classified as offensive. $m(r_i)$ is then the probability that the predicted rationales r_i alone are classified as offensive, and $m(x_i \backslash r_i)$ is the probability of the sentence with removed predicted rationales.

Now we can define **comprehensiveness** as

$$m(x_i) - m(x_i \backslash r_i) \tag{3.1}$$

. (Were all rationales needed to make a prediction?). The higher the score, the better. It can also become negative: then the model is more confident without rationales.

The metric **sufficiency** is defined as

$$m(x_i) - m(r_i) \tag{3.2}$$

. (Do extracted rationales contain enough signal?). The lower the score, the better.

For continuous rationales, the question of how to compute $x_i \backslash r_i$, in other words how to remove the rationales, arises. This is because here, every word gets a weight measuring how likely the word is predicted as a rationale. ERASER defines to remove the top k rationales again with a threshold, but we do not need this for applying the metrics to rule-based systems.

3.1.3 Application: predicting rationales with rules

One or more POTATO rules can be matched on the sentence which has to be classified. If e.g. the one-node rule "hate" matches on "I hate people", then this sentence is classified as e.g. offensive. A special property of ERASER is the fact that it cannot calculate samples with empty rationale ground truth. To solve this, every non-offensive HateXplain ground truth data was discarded before metric calculation (as done in the original HateXplain work).

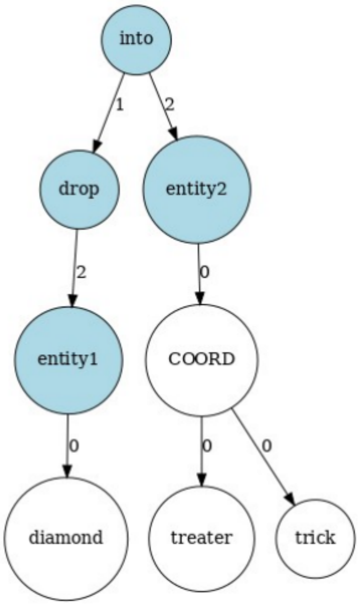


Figure 3.1: Example sentence graph in POTATO. If a rule matches, the words of the matching subgraph (blue nodes) is returned as the list of predicted rationales, so in this example "into", "drop", "entity1" and "entity2" [Kov22].

Plausibility Now in order to calculate the first metric, one first has to think about how POTATO is able to predict rationales. The idea here is simple; the matched subgraph is returned as the rationale. With our one-node rule example, "hate" is returned as the only rationale of the sentence. Other heuristics are also possible. Figure 3.1 shows another example of a given sentence and a matching subgraph including the resulting rationale prediction. We also lemmatize the ground truth rationales as works better with comparing the predicted rationales of the rules, as they are also lemmatized. With this, the two hard, discrete plausibility scores IOU F1 and Token F1 can be calculated.

Faithfulness The original probability function $m(x)$ is actually a continuous function between 0 and 1, as deep learning logit output is usually continuous. However, a POTATO rule either matches fully or not at all. Therefore, single sentence faithfulness metrics can be either 0 or 1, which is no problem. By aggregating multiple sentences, the faithfulness metrics are smoothed out. Therefore, to calculate all different $m(x)$ values needed (normal sentence - $m(x_i)$, rationales only - $m(r_i)$, sentence without rationales - $m(x_i \backslash r_i)$), one just has to classify three times with the given input data, convert to ERASER format and the framework calculates the comprehensiveness and sufficiency metrics by the given subtraction formulas for those two (See Equations (3.1) and (3.2)).

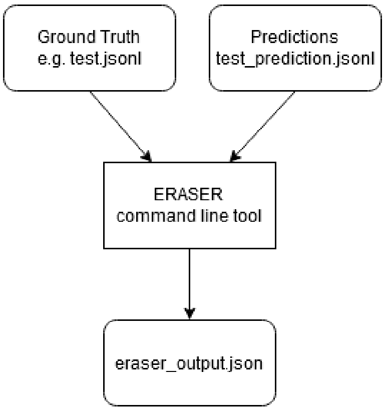


Figure 3.2: Model classification and rationale prediction and ground truth have to be converted into the ERASER jsonl format. After that, ERASER can be called and outputs a json object containing all the metrics.

3.1.4 ERASER workflow

After the model prediction is finished, the output has to be converted into ERASER format to start metric calculation. ERASER needs two different kind of input files; one for the ground truth and one for the prediction. Then, ERASER can be called and outputs a json file with their metrics as well as standard metrics like precision and recall (see Figure 3.2).

The files have to be in jsonl format, so every line has a json object containing a data sample. The formats between ground truth and prediction are slightly different, the documentation about how to encode the data can be found on their website [DJR+].

3.2 Preprocessing the data

While the HateXplain baseline models solely concentrate on the label classification problem [MSY+21], we want to look at the target annotation with our models. "Women" and "Homosexual" were chosen as the targets we want to study further in this work. After an initial test of the rules on these, our qualitative error analysis showed that human annotation contains overlapping targets, so there are annotations where e.g. two voters say the post is against black and one say it is against women. There is also the option for an annotator to state multiple target groups. Therefore, we introduce our own datasets derived from HateXplain. Each data point is annotated by at most 3 annotators. We apply both a purity filter as well as two different kinds of voting to create different dataset variants for the Women and Homosexual target. Regarding the resulting rationales, we use the same method as HateXplain and take the union of all rationales. No further preprocessing like removing the numbers and dates has to be done. Only the phrase "I'm" is exchanged with "I am" in order to improve the graph parsing of the rules.

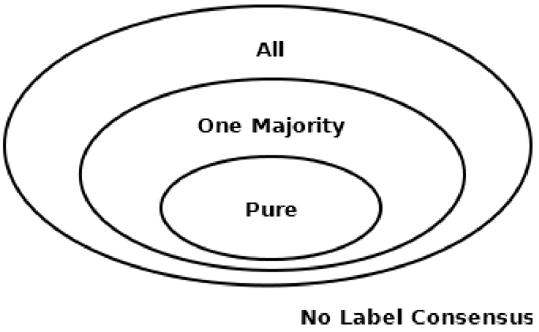


Figure 3.3: Visual representation of the different filters, namely "all", "one_majority" and "pure". Note that "all" does not include the data points where every annotator classified differently.

3.2.1 Purity filter

First of all, the purity filter represents 3 levels of purity to avoid the noisiness of the target annotations. The subsets are "all", "one_majority" and "pure", as seen in Figure 3.3. The filter "all" is the default, removing only the posts where no label consensus is possible. "one_majority" means that the majority category is only one target (number of majority_targets is length 1), and "pure" says there was only one type of target (the target_list is either length 1 or length 2, but one of those is a None). On other words, we count the number of targets with majority and the number of not-"None" targets per post and remove it if one of those is bigger than 1 according to our filter. Table 3.1 shows examples which annotations are inside which purity set. Note that an annotator can choose multiple targets.

Table 3.1 shows several examples of different purity.

Annotations (anno1, anno2, anno3)	**All**	**One Majority**	**Pure**
["African"], ["Women"], ["Homosexual"]	False	False	False
["Women","African"], ["None"], ["Homosexual"]	False	False	False
["Jewish","Homosexual"], ["Jewish","Homosexual"], ["Jewish","Homosexual"]	True	False	False
["Homosexual","Women"], ["Homosexual"], ["Women"]	True	False	False
["Homosexual","African"], ["African"], ["African"]	True	True	False
["None"], ["Women"], ["Women","Homosexual"]	True	True	False
["Women"], ["Women"], ["Women"]	True	True	True
["Women"], ["None"], ["Women"]	True	True	True

Table 3.1: Examples of different target annotation combinations and their purity.

3.2.2 Voting

Due to 3 targets per post, we need to decide the final class, which is in case of the women dataset either "Women" or "None", or in case of the homosexual dataset "Homosexual" or "None". We therefore introduce the two different votings "majority" and "minority". Majority voting means that at least 2 annotators voted for the target, whereas minority voting means that at least 1 annotator voted for the target. Examples are shown in Table 3.2.

Annotations (anno1, anno2, anno3)	**Target**	**Voting**	**Resulting label**
["None"], ["None"], ["None"]	Women	majority	not offensive
["Women"], ["None"], ["None"]	Women	majority	not offensive
["Women"], ["Women"], ["None"]	Women	majority	offensive
["Women","Homosexual"], ["None"], ["None"]	Homosexual	majority	not offensive
["Women","Homosexual"], ["None"], ["None"]	Women	majority	not offensive
["Women","Homosexual"], ["Women"], ["None"]	Women	majority	offensive
["Women"], ["Women"], ["Women"]	Women	majority	offensive
["None"], ["None"], ["None"]	Women	minority	not offensive
["Women"], ["None"], ["None"]	Women	minority	offensive
["Women"], ["Women"], ["None"]	Women	minority	offensive
["Women","Homosexual"], ["None"], ["None"]	Homosexual	minority	offensive
["Women","Homosexual"], ["None"], ["None"]	Women	minority	offensive
["Women","Homosexual"], ["Women"], ["None"]	Women	minority	offensive
["Women"], ["Women"], ["Women"]	Women	minority	offensive

Table 3.2: Examples of different target annotation combinations and the resulting label with either majority or minority voting.

3.2.3 Nomenclature

Finally, we can generate 2*3*3=18 files for each target with the naming convention being the following:

{majority/minority}_{train/val/test}_{all/one_majority/pure}

generated for either women or homosexual.

The first variable is the voting, if it was a majority, or any ("minority"). The second is the train/validation/test split, which was already given. The term "combination", which might show up later, means that all three splits are combined in one file. The third is the "purity", so pure means that there was only one type of target, one_majority if the majority category is only one target and all being all the data.

3.2.4 Statistics

We now present some basic statistics to get an overview of the differences between the datasets. We calculate the average and median number of texts, words, words in an offensive post, and rationales by annotators and union. Train/test/val splits and both

voting	dataset: all	# of texts	# of words		# of words in offn.		# of rationales							
			mean	med.	mean	med.	anno1		anno2		anno3		union	
							mean	med.	mean	med.	mean	med.	mean	med.
majority	train	15383	23.47	21	23.6	21	5.57	3	6.22	3	1.14	0	9.15	5
	test	1924	23.14	20	23.73	21	5.62	3	5.87	3	1.12	0	8.9	5
	val	1922	23.46	20.5	23.6	21	5.57	3	6.5	3	1.09	0	9.26	6
minority	train	15383	23.47	21	23.71	21	5.57	3	6.22	3	1.14	0	9.15	5
	test	1924	23.14	20	23.68	21	5.62	3	5.87	3	1.12	0	8.9	5
	val	1922	23.46	20.5	23.64	21	5.57	3	6.5	3	1.09	0	9.26	6

Table 3.3: Statistics of the original dataset without a specific target. As target, all HateXplain targets except "None" were considered.

voting	dataset: women all	# of texts	# of words		# of words in offn.		# of rationales							
			mean	med.	mean	med.	anno1		anno2		anno3		union	
							mean	med.	mean	med.	mean	med.	mean	med.
majority	train	15383	23.47	21	21.47	18	5.57	3	6.22	3	1.14	0	9.15	5
	test	1924	23.14	20	21.52	18	5.56	3	5.87	3	1.12	0	8.9	5
	val	1922	23.46	20.5	22.05	19	5.57	3	6.5	3	1.09	0	9.26	6
minority	train	15383	23.47	21	23.42	21	5.57	3	6.22	3	1.14	0	9.15	5
	test	1924	23.14	20	22.77	20	5.56	3	5.87	3	1.12	0	8.9	5
	val	1922	23.46	20.5	23.65	21	5.57	3	6.5	3	1.09	0	9.26	6

Table 3.4: Statistics of the women datasets with the "all" purity filter.

voting	dataset: women one_maj	# of texts	# of words		# of words in offn.		# of rationales							
			mean	med.	mean	med.	anno1		anno2		anno3		union	
							mean	med.	mean	med.	mean	med.	mean	med.
majority	train	12803	22.83	20	19.32	16	5.02	2	5.63	2	1.08	0	8.27	5
	test	1598	22.43	19	18.83	17	5.09	2	5.33	2	1.08	0	8.13	4
	val	1596	22.88	20	18.48	15	5.08	2	5.81	2	1.03	0	8.4	5
minority	train	12803	22.83	20	19.32	16	5.02	2	5.63	2	1.08	0	8.27	5
	test	1598	22.43	19	18.83	17	5.09	2	5.33	2	1.08	0	8.13	4
	val	1596	22.88	20	18.48	15	5.08	2	5.81	2	1.03	0	8.4	5

Table 3.5: Statistics of the women datasets with the "one_majority" purity filter.

voting types are compared in the same table. We do this for both targets (Women: Table 3.4, Table 3.5 and Table 3.6, Homosexual: Table 3.7, Table 3.8 and Table 3.9) with every purity filter and also for the original dataset (Table 3.3) for comparison, resulting in 2*3+1=7 tables.

We can see that the "all" purity filter indeed has the same number of elements as the original dataset. The stricter the filter, the fewer posts are in the set. Note that the voting type (majority or minority) only change the number of words in an offensive texts, which makes sense because voting only affects which posts are seen as offensive. We can also see that there is often no third annotation in the rationales. The median number of union rationales is 5, which is the reason the HateXplain deep learning models heuristically always predict 5 rationales.

3.3 Supervised learning

As already mentioned, a deep learning baseline should be compared to the rules. State of the art (see Section 2) mostly suggests a BERT-based model, which was, among other

voting	dataset: women pure	# of texts	# of words		# of words in offn.		# of rationales							
			mean	med.	mean	med.	anno1		anno2		anno3		union	
							mean	med.	mean	med.	mean	med.	mean	med.
majority	train	9296	21.77	18	17.61	14	4.34	2	4.78	2	1	0	7.04	4
	test	1174	21.56	18	16.08	12	4.08	2	4.41	2	1.01	0	6.61	3
	val	1146	21.54	18	15.29	12	4.21	2	4.95	2	0.93	0	6.96	4
minority	train	9296	21.77	18	18.32	15	4.34	2	4.78	2	1	0	7.04	4
	test	1174	21.56	18	17.09	15	4.08	2	4.41	2	1.01	0	6.61	3
	val	1146	21.54	18	15.79	13	4.21	2	4.95	2	0.93	0	6.96	4

Table 3.6: Statistics of the women datasets with the "pure" purity filter.

voting	dataset: homos. all	# of texts	# of words		# of words in offn.		# of rationales							
			mean	med.	mean	med.	anno1		anno2		anno3		union	
							mean	med.	mean	med.	mean	med.	mean	med.
majority	train	15383	23.47	21	21.8	19	5.57	3	6.22	3	1.14	0	9.15	5
	test	1924	23.14	20	19.35	16	5.62	3	5.87	3	1.12	0	8.9	5
	val	1922	23.46	20.5	20.79	17	5.57	3	6.5	3	1.09	0	9.26	6
minority	train	15383	23.47	21	22.98	20	5.57	3	6.22	3	1.14	0	9.15	5
	test	1924	23.14	20	20.92	18	5.62	3	5.87	3	1.12	0	8.9	5
	val	1922	23.46	20.5	22.83	20	5.57	3	6.5	3	1.09	0	9.26	6

Table 3.7: Statistics of the homosexual datasets with the "all" purity filter.

voting	dataset: homos. one_maj	# of texts	# of words		# of words in offn.		# of rationales							
			mean	med.	mean	med.	anno1		anno2		anno3		union	
							mean	med.	mean	med.	mean	med.	mean	med.
majority	train	12803	22.83	20	19.62	16	5.02	2	5.63	2	1.08	0	8.27	5
	test	1598	22.43	19	17.37	13	5.09	2	5.33	2	1.08	0	8.13	4
	val	1596	22.88	20	19.7	15	5.08	2	5.81	2	1.03	0	8.4	5
minority	train	12803	22.83	20	19.62	16	5.02	2	5.63	2	1.08	0	8.27	5
	test	1598	22.43	19	17.37	13	5.09	2	5.33	2	1.08	0	8.13	4
	val	1596	22.88	20	19.7	15	5.08	2	5.81	2	1.03	0	8.4	5

Table 3.8: Statistics of the homosexual datasets with the "one_majority" purity filter.

voting	dataset: homos. pure	# of texts	# of words		# of words in offn.		# of rationales							
			mean	med.	mean	med.	anno1		anno2		anno3		union	
							mean	med.	mean	med.	mean	med.	mean	med.
majority	train	9296	21.77	18	17.47	14	4.34	2	4.78	2	1	0	7.04	4
	test	1174	21.56	18	15.24	12	4.08	2	4.41	2	1.01	0	6.61	3
	val	1146	21.56	18	16.38	13	4.21	2	4.95	2	0.93	0	6.96	4
minority	train	9296	21.77	18	19.04	16	4.34	2	4.78	2	1	0	7.04	4
	test	1174	21.56	18	17.27	14	4.08	2	4.41	2	1.01	0	6.61	3
	val	1146	21.54	18	18	14	4.21	2	4.95	2	0.93	0	6.96	4

Table 3.9: Statistics of the homosexual datasets with the "pure" purity filter.

models, also used by HateXplain. The model code of HateXplain is open-source, we therefore adapt some of these models for comparison.

In total, there are 10 different runs reported in Table 5 in the HateXplain paper [MSY$^+$21]. Four different models are trained; CNN-GRU. BiRNN, BiRNN with an attention layer (BiRNN-Attn) and BERT. They were trained and evaluated with the 3-class dataset (offensive, hate, neutral). There are two different ways to train them: using only class labels and using the ground truth attention and class labels. The second option is only possible for models using attention, so not for BiRNN and CNN-GRU. The second option is called "<modelname>-HateXplain" in the their table, and will also be called by us this way. This option also has an additional lambda hyper-parameter controlling the weight of the additional attention loss. There are two methods to predict the rationales, LIME and attention. Again, attention is only possible with BERT and BiRNN-Attn. This results in the 2*1+2*2*2=10 runs; (CNN-GRU and BiRNN with LIME only, and BiRNN-Attn and BERT, trained two different ways and predicting the rationales two different ways)

By inspecting their results, we decide to use 3 models from HateXplain, namely **BERT-HateXplain** for its performance and **BiRNN-HateXplain** and **CNN-GRU** because they have the lowest sufficiency scores. BERT and BiRNN use the attention-based rationale prediction, while CNN-GRU uses LIME. To directly compare them to the rules, we have to adapt these models from the 3-class to the 2-class case, which the HateXplain authors already implemented partially. Further, we have to implement our two voting methods, as the HateXplain models need the combined dataset as input. We train with lambda=100, due to "Optimum performance occurs with λ being set to 100 for BiRNN with attention and BERT with attention in the supervised setting" [MSY$^+$21, p.7]. Training has to be done for each target (women, homosexual) and each voting (majority, minority). The models are programmed to always return the top 5 rationales.

3.4 Rule-based classification

One ruleset was developed for the "Women" and one for the "Homosexual" target class by using the HITL approach of POTATO and looking at the training data. The same ruleset is used for both voting types, so in total we just have 2 rule systems. The above discussed adaptions for predicting rationales with rules had no influence in creating the rules. The rule systems were created like any other POTATO rules. A separate evaluation script was created to predict and measure both the performance and explainability metrics.

For text parsing, UD is used. Rule and text graphs in POTATO can be represented by the so called PENMAN[1] notation by Goodman and Wayne [Goo20]. A single-node rule matching on the word "dyke" from the homosexual ruleset would look like the following:

```
(u_1 / dyke)
```

[1] The PENMAN notation is documented under https://penman.readthedocs.io/en/latest/notation.html.

This would not only match on every parsed text graph containing the word node "dyke", but also return "dyke" as a predicted rationale, as discussed in Section 3.1.3. Edges[2] can be formulated with a doublepoint; e.g.

```
(u_6 / dyke :nsubj (u_7 / I | we))
```

represents the two-node rule between the words "dyke" used as a nominal subject for either "I" or "we".

3.5 Qualitative methods

For inspecting the rules and its errors for the qualitative analysis, an inspection tool was created. For inspecting the rationale prediction for both the rules and the deep learning models, ERASER was adapted to output annotated and predicted rationales together with the IOU and Token scores.

For errors, we looked at the false positive and false negative classifications, to learn more about both the different systems predictions properties as well as about the target and rationale annotation. Classification, rationales and matched rules in case of the rule systems were inspected.

[2] Different edge labels and how they are parsed are documented on the universal dependency website `https://universaldependencies.org/en/dep/index.html`.

CHAPTER 4

Quantitative Results

After a short description of the experimental setup is given, we now present all of our numerical findings and discuss interesting facts about the quantitative results.

4.1 Experimental setup

First, we use the ERASER plausibility metrics to compare each annotator to the union in order to understand what maximum plausibility scores humans can reach. After that, we run our models on the data. In total, 3 deep learning models (CNN-GRU, BiRNN-HateXplain and BERT-HateXplain) and 1 rule system are tested per dataset. This is done for targets women and homosexual, with both majority and minority voting. For this thesis, we chose to first run on the "all" purity filter, and leave "one_majority" and "pure" for future work. Performance is calculated on the whole set of predictions, reporting precision, recall and F1 score. Also, we show how many posts were predicted as offensive and how many have the gold label offensive ("Homosexual" or "Women") per dataset. For explainability, we report the IOU, Token F1, precision and recall scores, and the both faithfulness metrics comprehensiveness and sufficiency.

4.2 Human performance on explainability

To compare the model results to human performance, we extracted the different annotator rationales and used ERASER to compare them on the union rationales. We again also report the results for the original dataset without a specific target group (just called "all"). Table 4.1 shows the annotator-union comparison for the original HateXplain dataset with our voting methods, while Table 4.2, Table 4.3 and Table 4.4 show them for our women datasets. Table 4.5, Table 4.6 and Table 4.7 show the annotator-union comparison for our homosexual datasets.

voting	dataset: all	IOU			Token		
		F1	P	R	F1	P	R
majority	anno1	0.827	0.847	0.807	0.84	0.954	0.818
	anno2	0.712	0.731	0.694	0.748	0.902	0.737
	anno3	0.491	0.351	0.819	0.348	0.383	0.354
minority	anno1	0.815	0.835	0.796	0.83	0.949	0.809
	anno2	0.729	0.748	0.711	0.762	0.915	0.745
	anno3	0.572	0.438	0.824	0.435	0.478	0.442

Table 4.1: Plausibility comparisons for single annotators against the union. This table shows the results for the original dataset, where any target except "None" is seen as offensive.

voting	dataset: women all	IOU			Token		
		F1	P	R	F1	P	R
majority	anno1	0.841	0.864	0.819	0.855	0.955	0.83
	anno2	0.736	0.759	0.714	0.767	0.895	0.757
	anno3	0.498	0.357	0.823	0.352	0.386	0.354
minority	anno1	0.838	0.86	0.817	0.852	0.957	0.829
	anno2	0.686	0.709	0.664	0.728	0.885	0.721
	anno3	0.404	0.271	0.795	0.269	0.298	0.273

Table 4.2: Annotator plausibility comparison for the women target and the purity filter "all".

The plausibility metrics can also be seen as a measure of similiarity. Comparing each annotator to itself yields a perfect 1. What we can observe is that the rationales from annotator 1 and 2 seem to be more similar to the union than from annotator 3. This makes sense, as annotator 1 and 2 only have 2-3 empty rationale annotations, while annotator 3 annotated 5605 toxic posts without rationales. This explains why the IOU F1 of annotator 1 and 2 is above 0.7, while for annotator 3, its around 0.5. We can therefore assume human plausibility performance lies around an IOU F1 of 0.7 and a little bit below for Token F1. The voting seems to not significantly change the results. Purer filtered data however seems to increase the scores a little bit.

4.3 Model performance

We now present the results of our model runs.

voting	dataset: women	IOU			Token		
	one_maj	F1	P	R	F1	P	R
majority	anno1	0.869	0.895	0.844	0.87	0.958	0.845
	anno2	0.77	0.795	0.747	0.792	0.908	0.78
	anno3	0.44	0.297	0.856	0.293	0.319	0.291
minority	anno1	0.839	0.864	0.816	0.849	0.955	0.824
	anno2	0.713	0.738	0.69	0.748	0.903	0.734
	anno3	0.388	0.254	0.818	0.251	0.277	0.254

Table 4.3: Annotator plausibility comparison for the women target and the purity filter "one_majority".

voting	dataset: women	IOU			Token		
	pure	F1	P	R	F1	P	R
majority	anno1	0.878	0.903	0.854	0.876	0.953	0.855
	anno2	0.817	0.841	0.794	0.829	0.927	0.813
	anno3	0.484	0.331	0.904	0.332	0.356	0.328
minority	anno1	0.887	0.911	0.864	0.886	0.957	0.866
	anno2	0.804	0.826	0.784	0.82	0.912	0.816
	anno3	0.436	0.287	0.904	0.288	0.308	0.285

Table 4.4: Annotator plausibility comparison for the women target and the purity filter "pure".

4.3.1 Model performance with explainability on true positives and false negatives

We first show the explainability results by using the same method as HateXplain, namely discarding the true negatives and false negatives in the ERASER calculation, as they contain empty rationales and cannot be used for explainability calculation. Performance is calculated on the whole prediction.

Model results for the women target are shown in Table 4.8 and Table 4.9. Homosexual target results are shown in Table 4.10 and Table 4.11. When comparing the rules to the deep learning systems, we learn that the former have the best precision, while the latter have the better F1 score, BERT-HateXplain having the best. The women ruleset does not perform that well in Recall, while the homosexual dataset nearly has a competitive F1 score.

If we look at the plausibility results, we can see that rule systems have a higher IOU F1 score, while deep learning models have a higher Token F1. This property is inspected further in the qualitative analysis. Looking at faithfulness, we see that the women ruleset is not very comprehensive, but both rulesets have quite good sufficiency scores. Still,

voting	dataset: homos. all	IOU			Token		
		F1	P	R	F1	P	R
majority	anno1	0.814	0.835	0.795	0.836	0.952	0.813
	anno2	0.736	0.751	0.721	0.776	0.918	0.76
	anno3	0.491	0.354	0.801	0.361	0.391	0.368
minority	anno1	0.821	0.84	0.803	0.842	0.955	0.819
	anno2	0.715	0.728	0.702	0.759	0.905	0.748
	anno3	0.476	0.34	0.792	0.342	0.376	0.347

Table 4.5: Annotator plausibility comparison for the homosexual target and the purity filter "all".

voting	dataset: homos. one_maj	IOU			Token		
		F1	P	R	F1	P	R
majority	anno1	0.818	0.842	0.795	0.833	0.957	0.807
	anno2	0.759	0.773	0.746	0.792	0.929	0.775
	anno3	0.463	0.325	0.808	0.334	0.359	0.342
minority	anno1	0.821	0.844	0.799	0.835	0.957	0.809
	anno2	0.741	0.754	0.729	0.777	0.919	0.762
	anno3	0.459	0.323	0.795	0.325	0.356	0.331

Table 4.6: Annotator plausibility comparison for the homosexual target and the purity filter "one_majority".

the BiRNN-HateXplain and CNN-GRU have the best sufficiency scores, being mostly negative, which was also the reason they were chosen as a comparison. Regarding the voting, one can see from the number of predicted gold labels that the minority datasets indeed contain more "offensive" labels than their majority counterpart. With minority voting, the rules have higher precision but lower recall and lower explainability. The deep learning models mostly get better performance with minority voting (note that the ruleset is the same for both voting types, while the deep learning models exist in two versions).

4.3.2 Model performance with explainability on true positives only

The explainability results presented before also contain false negatives. This means that predictions are incorporated into the score that do not contain rationales at all. Due to the rules being low-recall, we consider this to distort the results. Therefore, we additionally calculated the metrics on the true positives only. Note that the performance metrics were not calculated differently and are shown for the sake of completeness.

Tables 4.12 and 4.13 contain the true-positive (-_tp) only explainability results of the women target and Tables 4.14 and 4.15 contain the homosexual target results, all using

voting	dataset: homos. pure	IOU			Token		
		F1	P	R	F1	P	R
majority	anno1	0.836	0.856	0.817	0.849	0.957	0.827
	anno2	0.8	0.813	0.788	0.827	0.947	0.809
	anno3	0.508	0.367	0.828	0.379	0.41	0.386
minority	anno1	0.84	0.859	0.822	0.853	0.958	0.823
	anno2	0.791	0.804	0.779	0.822	0.942	0.808
	anno3	0.499	0.357	0.828	0.369	0.398	0.375

Table 4.7: Annotator plausibility comparison for the homosexual target and the purity filter "pure".

Dataset: women majority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	59/147	0.610	0.245	0.350	0.186	0.538	0.112	0.135	0.191	0.119	0.245	0.048
BERT-HateXplain	165/147	0.558	0.626	0.590	0.229	0.136	0.715	0.440	0.433	0.601	0.488	0.218
BiRNN-HateXplain	172/147	0.506	0.592	0.546	0.241	0.144	0.731	0.342	0.31	0.591	0.292	0.065
CNN-GRU	250/147	0.396	0.674	0.497	0.177	0.106	0.535	0.249	0.233	0.435	0.493	-0.086

Table 4.8: Performance and explainability metrics for the women target with majority voting and the "all" purity filter.

Dataset: women minority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	59/357	0.814	0.135	0.231	0.122	0.486	0.070	0.078	0.117	0.070	0.134	0.028
BERT-HateXplain	438/357	0.543	0.667	0.599	0.177	0.104	0.603	0.393	0.410	0.500	0.387	0.066
BiRNN-HateXplain	372/357	0.524	0.546	0.535	0.204	0.120	0.667	0.285	0.264	0.475	0.136	-0.048
CNN-GRU	391/357	0.494	0.541	0.516	0.140	0.082	0.466	0.201	0.187	0.333	0.369	-0.092

Table 4.9: Performance and explainability metrics for the women target with minority voting and the "all" purity filter.

purity filter "all". It is important to note that in this case, the comprehensiveness scores of the rule systems are perfect for the women target and nearly perfect for the homosexual target. One can also notice that the sufficiency of the women rules with minority voting are worst, especially for the women rules. The homosexual sufficiency score is comparable to the others. The Token F1 of the rules are now competitive to the deep learning models.

Dataset: homos. majority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	135/182	0.822	0.610	0.700	0.490	0.743	0.365	0.373	0.528	0.329	0.599	0.022
BERT-HateXplain	220/182	0.759	0.918	0.831	0.173	0.107	0.465	0.531	0.520	0.681	0.789	0.058
BiRNN-HateXplain	237/182	0.688	0.896	0.778	0.272	0.165	0.767	0.349	0.287	0.667	0.559	-0.017
CNN-GRU	220/182	0.727	0.879	0.796	0.247	0.150	0.700	0.311	0.252	0.605	0.694	-0.026

Table 4.10: Performance and explainability metrics for the homosexual target with majority voting and the "all" purity filter.

Dataset: homos. minority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	135/249	0.978	0.530	0.688	0.416	0.657	0.305	0.300	0.430	0.263	0.522	0.016
BERT-HateXplain	256/249	0.875	0.900	0.887	0.147	0.089	0.424	0.469	0.477	0.583	0.784	0.027
BiRNN-HateXplain	235/249	0.860	0.811	0.835	0.242	0.146	0.715	0.321	0.282	0.574	0.545	0.008
CNN-GRU	212/249	0.891	0.759	0.820	0.207	0.124	0.619	0.259	0.227	0.484	0.600	-0.004

Table 4.11: Performance and explainability metrics for the homosexual target with minority voting and the "all" purity filter.

Dataset: women_tp majority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	59/147	0.610	0.245	0.350	0.487	0.538	0.445	0.493	0.696	0.435	1.000	0.194
BERT-HateXplain	165/147	0.558	0.626	0.590	0.231	0.137	0.739	0.466	0.459	0.618	0.752	0.204
BiRNN-HateXplain	172/147	0.506	0.592	0.546	0.268	0.160	0.837	0.342	0.289	0.617	0.572	-0.093
CNN-GRU	250/147	0.396	0.674	0.497	0.200	0.119	0.612	0.264	0.248	0.478	0.682	-0.088

Table 4.12: Performance and explainability metrics for the women target with majority voting and the "all" purity filter. Explainability is calculated for true positives only.

Dataset: women_tp minority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	59/357	0.814	0.135	0.231	0.463	0.486	0.442	0.410	0.615	0.366	1.000	0.208
BERT-HateXplain	438/357	0.543	0.667	0.599	0.199	0.118	0.623	0.431	0.435	0.555	0.585	0.098
BiRNN-HateXplain	372/357	0.524	0.546	0.535	0.234	0.139	0.729	0.313	0.279	0.544	0.268	-0.147
CNN-GRU	391/357	0.494	0.541	0.516	0.181	0.107	0.597	0.241	0.218	0.417	0.495	-0.095

Table 4.13: Performance and explainability metrics for the women target with minority voting and the "all" purity filter. Explainability is calculated for true positives only.

Dataset: homos._tp majority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	135/182	0.822	0.610	0.700	0.643	0.743	0.567	0.563	0.798	0.498	0.982	0.036
BERT-HateXplain	220/182	0.759	0.918	0.831	0.181	0.111	0.476	0.553	0.540	0.706	0.866	0.043
BiRNN-HateXplain	237/182	0.688	0.896	0.778	0.280	0.171	0.783	0.354	0.291	0.684	0.632	-0.051
CNN-GRU	220/182	0.727	0.879	0.796	0.276	0.168	0.776	0.328	0.260	0.659	0.775	-0.025

Table 4.14: Performance and explainability metrics for the homosexual target with majority voting and the "all" purity filter. Explainability is calculated for true positives only.

Dataset: homos._tp minority_test_all	Performance				Explainability							
					Plausibility - IOU			Plausibility - Token			Faithfulness	
Model	Predicted	Prec.↑	Recall↑	F1↑	F1↑	Prec.↑	Recall↑	F1↑	Prec.↑	Recall↑	Comp.↑	Suff.↓
Rules	135/249	0.978	0.530	0.688	0.588	0.657	0.531	0.490	0.703	0.429	0.985	0.030
BERT-HateXplain	256/249	0.875	0.900	0.887	0.153	0.093	0.438	0.492	0.495	0.616	0.879	0.011
BiRNN-HateXplain	235/249	0.860	0.811	0.835	0.252	0.152	0.745	0.338	0.292	0.617	0.705	-0.062
CNN-GRU	212/249	0.892	0.759	0.820	0.244	0.147	0.706	0.291	0.245	0.569	0.747	0.019

Table 4.15: Performance and explainability metrics for the homosexual target with minority voting and the "all" purity filter. Explainability is calculated for true positives only.

dataset: women all	# of predicted rule rationales (tp&fp)		# of predicted rule rationales_tp	
	mean	med	mean	med
train	1.63	1	1.65	1
test	1.47	1	1.56	1
val	1.52	1	1.52	1

Table 4.16: Number of mean and median predicted rationales (tp & fp as well as tp only) of the women ruleset on the purity filter "all".

dataset: women one_maj	# of predicted rule rationales (tp&fp)		# of predicted rule rationales_tp	
	mean	med	mean	med
train	1.61	1	1.62	1
test	1.34	1	1.33	1
val	1.37	1	1.37	1

Table 4.17: Number of mean and median predicted rationales (tp & fp as well as tp only) of the women ruleset on the purity filter "one_majority".

4.3.3 Number of predicted rationales

Our deep learning models, taken directly from HateXplain, always predict 5 rationales, which they find the most likely. However, the rule systems do not generally return a fixed number of rationales. The number is rather dependent on which rules match, and there is the possibility that multiple rules may match. We therefore run the rules on every target, every purity filter, every data split and on majority voting to get a grasp on how many rationales our rules predict on average, and as the median. We do this with true positives and false positives, and with true positives only (called "_tp") in order to compare every non-empty rationale prediction (tp & fp) with the true positive-only case.

These statistics for the women rule system are shown in Table 4.16, 4.17 and 4.18, and for the homosexual rule system in Table 4.19, 4.20 and 4.21. We can see that in case a rule matches correctly, approximately 1 rationale is predicted on average. The number is a little bit lower for the homosexual ruleset and a little bit higher for the women target group.

dataset: women pure	# of predicted rule rationales (tp&fp)		# of predicted rule rationales_tp	
	mean	med	mean	med
train	1.52	1	1.58	1
test	1.35	1	1.4	1
val	1.28	1	1.3	1

Table 4.18: Number of mean and median predicted rationales (tp & fp as well as tp only) of the women ruleset on the purity filter "pure".

dataset: homos. all	# of predicted rule rationales (tp&fp)		# of predicted rule rationales_tp	
	mean	med	mean	med
train	1.17	1	1.17	1
test	1.13	1	1.14	1
val	1.16	1	1.14	1

Table 4.19: Number of mean and median predicted rationales (tp & fp as well as tp only) of the homosexual ruleset on the purity filter "all".

dataset: homos. one_maj	# of predicted rule rationales (tp&fp)		# of predicted rule rationales_tp	
	mean	med	mean	med
train	1.19	1	1.19	1
test	1.16	1	1.16	1
val	1.17	1	1.14	1

Table 4.20: Number of mean and median predicted rationales (tp & fp as well as tp only) of the homosexual ruleset on the purity filter "one_majority".

dataset: homos. pure	# of predicted rule rationales (tp&fp)		# of predicted rule rationales_tp	
	mean	med	mean	med
train	1.23	1	1.24	1
test	1.19	1	1.17	1
val	1.25	1	1.23	1

Table 4.21: Number of mean and median predicted rationales (tp & fp as well as tp only) of the homosexual ruleset on the purity filter "pure".

CHAPTER 5

Qualitative Results

In this chapter we present our qualitative findings by looking at different data samples and their classification / rationale prediction. The aim of this analysis is to find out why the rules perform the way they due in terms of performance and explainability.

5.1 Interpretation of performance

The first qualitative analysis was conducted before we introduced the concept of our purity filters. We however already had the majority and minority voting. Back then, we had a primary dataset, where at least 2 annotators agreed on the label, and a secondary dataset where at least 1 annotator labeled it as such. Both versions for target "Women" and "Homosexual". One can also say that the primary women dataset is the majority_women_all dataset, and the secondary women dataset is the minority_women_all dataset. The same can be said for the homosexual target datasets.

For all four validation sets, about 100 random samples were inspected with the inspection tool. Interesting errors and cases are discussed here. Although we concentrate on the classification errors in this Section, rationales are also commented. The analysis made us lemmatize the ground truth rationales and introduce the purity filters.

5.1.1 Women primary validation set

Table 5.1 shows the most interesting samples from the primary women dataset.

Looking at some false positives, WP-FP1 seems again to be an annotation error, as this annotation appears to be controversial. The same is the case for WP-FP2. WP-FP3 however shows how hard it is to use words like "slut" as a general offensive content indicator, as the author of this post most likely talks positively about herself. The same is most likely the case for WP-FP4, but with the word "bitch". WP-FP5 is an interesting case where the post is hate speech, but not against women.

ID	Text
WP-FP1	has stupid rich hoe
WP-FP2	<user> thot that hoe over there
WP-FP3	it spooky szn so if you want to send me money so i can buy spooky slut costumes please do", the author of this wants to buy "slut costumes
WP-FP4	gotta get you a real bitch a bad bitch ain t nothing but a bill
WP-FP5	<user> <user> fuck you die already you shit spewing pussy grabber i fucking hat$ you hope you choke on an egg mcmuffin
WP-FN1	all of my exes were cute but they were hoes i guess i only attract fly looking thots
WP-FN2	men must not confuse chivalry with letting women walk all over you
WP-FN3	evwimena says <date> at <time> what thought process do you expect from a sociali st snowflake dyke she has the support of the not sure if we are sitters or point ers restroom users when the o.b. delivered this young person the doc told her mo m she has a beautiful girl
WP-FN4	<user> <user> two <number> angel women
WP-FN5	just remember that the bitch commie girlies and their low iq muds are going to s link beneath the shield of the cuckservatives and alt lite when their independen t operation folds like a congolese coup then we will have our civil war with som eone who can fight

Table 5.1: Interesting samples for the women primary validation set.

Looking at some false negatives, WP-FN1 was classified as such because the rule "hoe" did not match on "hoes". WP-FN2 is a controversial annotation. For WP-FN3, the right rationale should have been "dyke", which might slightly overlap with the homosexual target. WP-FN4 might be an annotation mistake, as the context does not seem offensive. WP-FN5 is an interesting annotation, as it contains 18 ground truth rationales.

In general, several incidents were seen where "bitches" was the ground truth but the rule system returned "bitch", again due to lemmatization by the graph parser. What the ruleset does quite good from a subjective view is splitting women from other target groups. The rule author remarks that the profane words hoe/slut/bitch are not necessarily the best indication of offending women, even though these words are present in a lot of the true positive cases.

5.1.2 Women secondary validation set

The procedure was repeated for the secondary dataset. Table 5.2 shows secondary false positives and false negatives.

Due to the high precision of the rules, there are not many false positives available. WS-FP1 and WS-FP3 were already shown in the primary set. WS-FP2 is another example of seemingly wrong annotation. The same might be true for WS-FP4, but here, the main target annotation might not be women. WS-FP5 is an interesting post, which annotation's offensiveness is not clear to us if the post author calls the city Chicago a "broke bitch". (The simple negative usage of this word might already be.)

Regarding false negatives, we have posts from the primary set like WS-FN1, but also seemingly wrong target classifications like WS-FN2, which main target might rather be against homosexual. However, also mixture of targets exist, like e.g. WS-FN3, because

ID	Text
WS-FP1	has stupid rich hoe
WS-FP2	<user> <user> attacking me for my appearance doesn ' t work when i got hoes in t he inbox sending titties faggot
WS-FP3	it spooky szn so if you want to send me money so i can buy spooky slut costumes please do", the author of this wants to buy "slut costumes
WS-FP4	i too hate the current shitty times when black races are everywhere and the new anti white gestapo laws hunt us down i do not want to see white sluts and black nigs arabs couples on the street makes me sick
WS-FP5	paid my last speeding ticket today <unicode symbol> also chicago you can kiss my ass you broke] bitch stop taking my money <unicode symbol>
WS-FN1	all of my exes were cute but they were hoes i guess i only attract fly looking thots
WS-FN2	who the fuck you faggots think you are
WS-FN3	rosie is an insane lesbian being a lesbian tells you she is mentally ill
WS-FN4	defending muslims being in europe saying that islam isnt the greatest threat to the white race pick one kys for having the reading comprehension of a goat
WS-FN5	thick women are only fat women who fat is in places that men like fat

Table 5.2: Interesting samples for the women secondary validation set.

lesbians are homosexual women. Other false negative examples are WS-FN4 (against islamic people) and WS-FN5, which does not contain any ground truth rationale, but is still voted as against women.

General findings were again lemmatization issues and that here, plenty of "Women" posts had no ground truth rationales, and the profane word dike/dyke appears sometimes. The system again matches on "hoe" and "bitches" wrongly, but mostly, the annotation seems wrong.

5.1.3 Homosexual primary validation set

Table 5.3 shows selected samples for the primary homosexual dataset.

Looking at the primary homosexual false positives, one can again identify probable wrong annotations like HP-FP1, HP-FP3 and HP-FP5. HP-FP2 is an interesting case where the offended group is the LGTBQ community, the voting however did not find any Homosexual target in it and it was therefore voted to be non-offensive. HP-FP4 also seems to offend multiple target groups (black and homosexual people).

Seemingly wrong annotations also appear at false negatives, like HP-FN5. HP-FN1 is actually defending homosexual marriage, but it seems like the annotators considered this post as offensive. The content of HP-EN3 is not meant to be offensive against homosexual people, but it is very profane. Posts like HP-FN2 are not really expressive. This post in particular contains the single rationale "gay". Within HP-EN4, the rule containing "lesbian" did not match.

Again, problems with the lemmatization ("faggot" vs "faggots") occur. Other than that, one can see that in this dataset one-word rationales, specifically the words "faggot" and "dyke", are very common. If the ground truth however has more than two rationales, the

ID	Text
HP-FP1	will the cowardly little nazi faggot that downvoted u.s. marines kindly identify yourself so we can all mock you thank you
HP-FP2	there are just women and men male and female there is no gender spectrum the on$ s that claim they are such and such are just in need of help i e psychiatric th$ only ones i accept other than male and female are gay and even then some dabble in this
HP-FP3	biggest safe space online they banned me for being too pro trump and questionin$ beta faggots who had not sworn allegiance to the god emperor apparently i was $ oncern trolling please mods are pussies who are not willing to do what it takes to take their community to the next level
HP-FP4	bullshit every fucking liberal i have ever met assumes shit and lies about every thing you cannot be a liberal and a fucking true christian at the same time so g o fuck yourself nigger faggot liar your just another tool
HP-FP5	dont kid yourself you look like a cross dressing faggot himey boy now take a hik e kike
HP-FN1	and it fine n dandy if you wanna get married but that not going to raise up the plummeting marriage rates you will not get them up by calling guys gay or telli$g them to man up if you want traditional marriage go and make it traditional ag$in
HP-FN2	<user> hope his gay
HP-FN3	people on twitter will go fucking ballistic if you say something negative about their fat anime tits and say why are you judging me for what i like its wrong to judge people and then follow that up by calling you a faggot or something of th e sort
HP-FN4	rosie is an insane lesbian being a lesbian tells you she is mentally ill
HP-FN5	hi pls kiss me full homo

Table 5.3: Interesting samples for the homosexual primary validation set.

correct rationale set is never predicted. This might be because of the current heuristic to predict rationales from the subgraph of the rules, which means that the quality of rationale prediction depends on the rule structure. Good classifying and good rationale predicting rules might be opposing goals when building rule systems.

5.1.4 Homosexual secondary validation set

We lastly look at the secondary homosexual dataset and some samples shown in Table 5.4 and begin with false positives.

HS-FP2 shows that "faggots" is hard to classify. The problem with HS-FP4 is similar. Another question is indeed how to treat FP sentences like "i am the biggest faggot cunt" (HS-FP1). The rule systems do not have the context knowledge, but even if the rule author has them, it is hard to construct rules which reproduce the exact classification behaviour. HS-FP5 also needs context to be interpreted as non-offensive. HS-FP3 shows the ultimate profane post.

Regarding false negatives, we also saw HS-FN1 in the analysis of the primary homosexual dataset. The person is talking about a slip out of a tongue. We consider the annotation should be "None". HS-FN2 and HS-FN3 are another context examples, but we believe they are both rightfully labelled as offensive against gay relationships. HS-FN4 again appeared from the primary dataset. The last interesting post is HS-FN5, offending homosexuals decently with slight sarcasm.

ID	Text
HS-FP1	i am the biggest faggot cunt
HS-FP2	yeah im a faggot and what are you disrespectful id rather be a faggot than a disrespectful child any day
HS-FP3	over the last week i was thinking that altretard pedo nazi faggot would become $ike the ultimate insult and then pewdiepie said nigger and the world came to an end no no apparently even the feelings of the altleft are retarded and unimagin$taive
HS-FP4	i hope you all are aware that i will definitely keep watching queer eye for the serotonin and faggotry right
HS-FP5	do not call my phone with nothing to talk about this late at night and definite$ y do not call me phone and use slurs these slurs include retard faggot tranny sp ic chink and other offensive ass word i will literally never talk to you again
HS-FN1	i was talking to a co worker about the local takeaway shops and instead of saying maggots i said faggots and i am ready to go kill myself oh my god maddie
HS-FN2	<user> <user> just like close friendship is basicly gay relationship
HS-FN3	<number> of <number> being one court ruling away from yet another national moral failure is no way to live also what has scotus given us over the years but sanc tioned infanticide forced racial integration birthright citizenship for foreigne rs homosexual marriage the god given right to self determination being declared nonexistent
HS-FN4	rosie is an insane lesbian being a lesbian tells you she is mentally ill
HS-FN5	russell also writes about how jewish and italian mobsters founded las vegas jewi sh copyright thieves founded hollywood and gay mafiosos helped spark the gay rig hts movement he thinks all of these things were good for society of course

Table 5.4: Interesting samples for the homosexual secondary validation set.

Sometimes offensive words appear which are not in the rules yet (e.g. "kike" dyke). However, adding these might lead to overlearning on the dataset. It is quite rare that sentences attack the minority in a non-direct way, but those are hard to interpret by the rule system. Like with the women datasets, one can say that often, there is no ground truth rationale rationale even if the class is offensive.

5.2 Interpretation of plausibility

We now investigate the reason why for deep learning models, the Token F1 is higher, while for rule systems, it is the IOU F1. We already know that the deep learning models always predict 5 rationales. We learned from Section 3.2.4 that the union rationales have a median of 5, but looking manually at the ground truth shows that often, there are 1-3 rationales, hinting that the distribution is quite skewed. The rules however predict 1-2 rationales. The rule which is closest to predicting a rationale set with 3 or more words is shown with W-TP12 in Table 5.6. Other than that, We found that only 2-word rationales are matching perfectly at most.

Why is the IOU F1 smaller than the Token F1 for the deep learning models? By looking at random datapoints like in Table 5.5, we learned that the IOU F1 score is only more forgiving when little bit less or equal the number of ground truth rationales are predicted. The maximum possible IOU score can be written as a function of the predicted words x. Maximum IOU = $min(x, a)/max(x, a)$, were a is the ground truth

Ground truth	Prediction	IOU F1	Token F1
['pussy']	['pussy']	1.000	1.000
['bitch', 'bitch', 'shit', 'nigga']	['bitch', 'bitch']	1.000	0.667
['wetback', 'bitch']	['bitch']	1.000	0.667
['bitch']	['get', 'bitch', 'get']	1.000	0.333
['bitch']	['get', 'bitch', 'get']	1.000	0.333
['faggot', 'sand', 'nigger']	['faggot']	1.000	0.667
['kid', 'kike', 'spic', 'nigger', 'faggot', 'allah', 'a', 'fat', 'bitch']	['faggot']	0.000	0.000
['dykes']	['mdykes', 'they', 'was', 'like', 'i']	0.200	0.333
['moslem', 'fag']	['the', 'minister', 'fuck', 'fag', 'moslem']	0.400	0.571
['faggot', 'jew', 'or']	['faggot', 'jew', 'a', 'or', 'you']	0.000	0.745
['homosexual', 'and', 'with', 'i', 'me', 'i', 'for', 'guys', 'say', 'only', 'am', 'malfunction', 'dumb', 'so', 'and', 'because', 'women', 'things', 'hang', 'out', 'do', 'unbelievably', 'make', 'because']	['mi', 'homosexual', 'unbelievably', 'so', 'am']	0.000	0.357
['pussy', 'mutes', 'bitch', 'the', 'thanks', 'faggot', 'the', 'and', 'for', 'scalp']	['pussy', 'mutes', 'the', 'thanks', 'faggot']	0.000	0.667

Table 5.5: Example rationale ground truth and predictions and the resulting IOU and Token F1 scores from the validation sets. Note that duplicate rationales are actually from different parts of the sentence.

number of rationales. If x<a, the maximum IOU increases linearly till it reaches a. If x>a, the IOU decreases again with ~1/x.

This is what actually happens when the deep learning models "guess" 1 correct and 4 random rationales. The IOU denominator gets big, the IOU gets under 0.5 and then the whole term becomes 0 because of the threshold, whereas the Token F1 would still be 0.2. So the IOU punishes the deep learning models for always predicting 5 rationales.

5.3 Interpretation of faithfulness

We now observe the faithfulness metrics to learn why the initial assumption, that both comprehensiveness and sufficiency for rules are perfect, does not hold. When looking at the values of the faithfulness, there are two cases.

Assume the rule matches. Then, $m(x_i)$ equals 1, because the probability of the offensive class predicted by the rule is 100%. $m(x_i \backslash r_i)$ then must yield 0 nearly every time, as removing the rationales makes the rules not match anymore. $m(r_i)$ on the other hand might be 0 or 1, sometimes the rule matches on the newly parsed subgraph again, but not always. The resulting comprehensiveness must therefore nearly always be 1-0 = 1. (There seems to be one exception in the homosexual test set, as seen in Table 4.14.) The resulting sufficiency is mostly 1-1=0, but sometimes, it can be 1.

Assume the rule does not match. The post is then classified as "None". Then, $m(x_i)$ equals 0, because the probability of the post being offensive is predicted as 0%. Removing the rationales (There are none.) yields $m(x_i \backslash r_i) = 0$. Predicting only on the rationales (There are again none.) also yields $m(r_i) = 0$. The resulting comprehensiveness is therefore 0-0 = 0, and the resulting sufficiency mostly 0-0 = 0.

Why is comprehensiveness not a perfect 1? In the case where we input both true positives and false negatives to ERASER, averaging all comprehensivenesses gives something smaller 1 due to false negative "None" classifications being 0. This is right because actually everything ERASER consideres should be offensive and has rationales, and due to None being the wrong classification and having no rationales, $m(x_i) - m(x_i \backslash r_i)$=0. Note that comprehensiveness is actually 1:1 moving with the Recall: "In general, the rationales we have for tasks are sufficient to make judgments, but not necessarily comprehensive. However, for some datasets we have explicitly collected comprehensive rationales for at least a subset of the test set. Therefore, on these datasets recall evaluates comprehensiveness directly (it does so only noisily on other datasets). " [DJR+19, p. 5].

When looking at the homosexual all test set after the numbers were calculated, one can see the comprehensiveness score is not perfect due to one single case where one rationale is not being removed correctly. The original sentence is "q why did the bulldykes beat the faggots in a <number> mile run a because the dykes went lick ity split but the fags had to pack their shit" and $m(x_i \backslash r_i)$ is "q why did the bulls beat the s in a <number> mile run a because the s went lick ity split but the fags had to pack their shit", so the rationale "fags" should have been removed.

Why is sufficency not a perfect 0? Sufficency with offensive classifications is not always 0, because $m(r_i)$ does not always yield 1. This means that the rule was not able to match on the new graph parsed from the subgraph words alone. All samples where sufficency is 1 are shown for the women majority validation dataset in Table 5.6 and for the homosexual majority validation dataset in Table 5.7. Its interesting to see that in the cases of W-TP6, W-TP8 and W-TP10, a rule consisting of two non-connected nodes are matching, therefore only predicting one rationale of the first node.

ID	Text	Predicted rat.	Matched rules
W-TP1	told dat bitch get right or get left	[’get’, ’bitch’]	[’(u_1 / what|want|stay|t|make|get|pull|stupid|act :nsubj (u_0 / bitch))’]
W-TP2	found that psychopath guilty bitch want have to worry about texting gone get all the attention she needs them bull dykes waiting <unicode symbols>	[’want’, ’bitch’]	[’(u_1 / what|want|stay|t|make|get|pull|stupid|act :nsubj (u_0 / bitch))’]
W-TP3	i am really pissed off tho lmao i hate dirty bitches	[’bitch’, ’hate’]	[’(u_1 / tell|like|make|fuck|break|hate|trust :obj (u_0 / bitch))’]
W-TP4	you betta go get my money lil bitch i gotta stack whore i can go str8 for the flush or kiss it off the backboard <unicode symbol>	[’bitch’]	[’(u_0 / bitch)’, ’(u_1 / hoe|slut|whore)’]
W-TP5	infatuated wit pretty bitches	[’bitch’, ’pretty’]	[’(u_1 / bitch|whore|slut :amod (u_0 / nasty|fat|many|right|heartless|sexy|lie|fuck|dry|fake|disgusting|ugly|pretty|communist|break|racist|loving))’]
W-TP6	niggas is fucking pussy whole dick riding bitches do not come around here with that faggot shit you peon	[’pussy’]	[’(u_1 / pussy)’, ’(u_0 / bitch)’]
W-TP7	democrats from minnesota leftist muzzie shithole are the dumbest anti americans there are klobuchar one of the reps is a dumb bitch who has a moronic position on the tax bill fuck this stupid woman	[’woman’, ’stupid’]	[’(u_1 / woman|female|feminist|bitch :.* (u_0 / crazy|mad|stupid|undatable|ugly|liberal))’], [’(u_1 / woman|female|feminist|bitch :nsubj (u_2 / I))’]
W-TP8	<user> it do not matter bitch you still selling pussy trash ass how	[’pussy’]	[’(u_1 / pussy)’, ’(u_0 / bitch)’]
W-TP9	i despise it when news personalities like lauren southern say they are not refugees they are economic migrants no you dumb kike bitch they are i n v a d e r fuck it after what happened to pamela shoot em if they do not leave	[’bitch’, ’kike’]	[’(u_12 / bitch :compound (u_2 / b|kike|shit|bitch))’]
W-TP10	must thought i was pussy why you playin wit these bitches	[’pussy’]	[’(u_1 / pussy)’, ’(u_0 / bitch)’], [], ’Women’]
W-TP11	when i watched clinton play the woman card during the election it seemed pretty clear to me that this is why women can not be in politics the fact that everybody did not immediately conclude that is a mystery i do not even see how people rationalize it	[’not’, ’politics’, ’woman’, ’can’]	[’(u_0 / .* :nsubj (u_31 / woman) :aux (u_33 / can) :advmod (u_34 / not))’]
W-TP12	the redskins is trending is there us teams also called the niggers spics kikes bitches or faggots asking for a friend	[’bitch’, ’kike’]	[’(u_12 / bitch :compound (u_2 / b|kike|shit|bitch))’]

Table 5.6: All posts with a sufficency of 1 from the women majority_val_all dataset.

<table>
<tr><th>ID</th><th>Text</th><th>Predicted rat.</th><th>Matched rules</th></tr>
<tr><td>H-TP1</td><td>just putting this out here most queers are alcoholics <unicode symbol></td><td>['alcoholic', 'queer']</td><td>['(u_1 / .* :obj|nsubj (u_2251 / queer))'], ['(u_1 / call|love :obj (u_2251 / queer))', ' (u_2 / war|ship)']</td></tr>
<tr><td>H-TP2</td><td>as all of us know queers like to masturbate in the intestines of someone else obviously this does not mean they d like to do that moreso in the intestines of someone who is not yet of legal age or who cannot reproduce themoaryoulearn</td><td>['like', 'queer']</td><td>['(u_1 / .* :obj|nsubj (u_2251 / queer))'], ['(u_1 / call|love :obj (u_2251 / queer))', ' (u_2 / war|ship)']</td></tr>
<tr><td>H-TP3</td><td>figure skating ladies and pairs eff the queers</td><td>['eff', 'queer']</td><td>['(u_1 / .* :obj|nsubj (u_2251 / queer))'], ['(u_1 / call|love :obj (u_2251 / queer))', ' (u_2 / war|ship)']</td></tr>
<tr><td>H-TP4</td><td><user> frere t ' es gay</td><td>['user', 'gay']</td><td>['(u_1 / nigg.*|hate|group|parade|word|month|user|kind|kill |act|ass|one|picture :.* (u_2 / gay))'], ['(u_2 / gay :nsubj (u_4 / I))']</td></tr>
<tr><td>H-TP5</td><td>i refuse to become homosexual but sometimes i really fucking hate these african american hood rat nigger bitches <unicode symbol> goat</td><td>['homosexual']</td><td>['(u_5095 / homosexual|homo)', ' (u_1 / rape|pervert|sick|ill|insane|fuck|pedo.*)']</td></tr>
<tr><td>H-TP6</td><td>fuck your data you sound like a homo i am not going to get sucked into your gay game of name the fed for all i know you are the fed alex jones is definitely some kind of fed or jew loving gatekeeper or crypto</td><td>['homo']</td><td>['(u_5095 / homosexual|homo)', ' (u_1 / rape|pervert|sick|ill|insane|fuck|pedo.*)']</td></tr>
<tr><td>H-TP7</td><td>lol the ns germans are all gay they are too busy having sex with each other while their women are being raped <laugh></td><td>['gay', 'rape']</td><td>['4((u_1 / gay), (u_3 / fuck|shit|stupid|rape|death|pedophile|pedo|ill|sick|ugly |stupid|insane|depressed|ugly))']</td></tr>
<tr><td>H-TP8</td><td>glad to hear do not be a sheltered cuck for any side lefties are gay as fuck but gayness is contagious to the right as well but i think you already grasp that</td><td>['gay', 'fuck']</td><td>['4((u_1 / gay), (u_3 / fuck|shit|stupid|rape|death|pedophile|pedo|ill|sick|ugly |stupid|insane|depressed|ugly))']</td></tr>
<tr><td>H-TP9</td><td>it said that at some point during their life every man will have a homosexual fantasy mine is to kick the fuck out of one</td><td>['homosexual']</td><td>['(u_5095 / homosexual|homo)', ' (u_1 / rape|pervert|sick|ill|insane|fuck|pedo.*)']</td></tr>
</table>

Table 5.7: All posts with a sufficency of 1 from the homosexual majority_val_all dataset.

CHAPTER 6

Discussion

In the last chapter, we firstly discuss possible future work. Then, we conclude our results while giving a short summary.

6.1 Future work

For our upcoming paper, we will compute the metrics not only on the "all" filter but also experiment with the "one_majority" and "pure" datasets. These filters could also be inspected qualitatively. Further, the way we adapted our rules systems for predicting rationales and running the ERASER explainability metrics is just one way to achieve this. There are several parameters which can be tested in future experiments. For example, we did not compare rules and deep learning systems using the continuous AUPRC score, but this could be done by smoothing the hard rationale predictions with e.g. sigmoid functions. It is also just a heuristic to return all of the matching subgraph nodes of the rules as the rationales. This mechanic could also be changed, but it may come with the price of faithfulness, as it is currently easy to predict what words are returned as rationales by looking at the rules alone.

Speaking of faithfulness, we currently reparse every text after removing the rationales. However, this is just one way to interpret the formulas of $m(x)$ and $m(x\backslash r)$. Another interpretation would be to manipulate the graphs without reparsing the sentence. For example, instead of removing the rationale "faggot" from the sentence "You are a faggot", reparsing the graph and then matching the rules again on the resulting graph, which is currently done, one could remove the faggot node from the already parsed graph. However, it has to be discussed if this is fair for the rule systems, as this preserves more of the structure of the sentence graph. It might look different when being reparsed without the rationale words.

We have also learned that literature sees explainability metrics and faithfulness in general in multiple ways. The metrics we used to compare our models are not absolute, so one could also use different means of measurement, or for example the normalized ERASER metrics (where faithfulness scores are always between 0 and 1) for comparing and interpreting the numerical results better. Looking at the big number of target classes, it also has to be stated that we just studied a fraction of all possible target classes of HateXplain. More rule systems, supervised models and datasets can be created on the rest of the targets, e.g. Black, Arab, Refugee, etc.

6.2 Conclusion

Finally, to review our research goals, we did not only answer RQ 1 and 4 thoroughly in Section 4 and RQ 2 and 3 in Section 5. Several other aspects were observed while studying explainability on the example of HateXplain.

We introduced a way to preprocess the HateXplain dataset into datasets of different purity and target group. We found a way to implement rationale-prediction for rule systems and looked at the differences of predicting hate speech text between current state-of-the-art deep learning systems and rule systems in dimension of performance and explainability. Rules have higher precision, Deep learning models have higher F1 scores. We learned that the two tasks of detecting offensive content against women and homosexual people are two different kind of tasks, and that well-performing rules do not automatically yield well explaining rationales. We also question the gold-label status of human rationale annotations and show this by comparing human performance. Finally, we show the nearly perfect comprehensiveness and good sufficency of rule systems when measured on true positives only, which lies in the very nature of how these systems work.

List of Figures

List of Tables

Acronyms

AI artificial intelligence. 3, 11

AMR Abstract Meaning Representation. 10, 11

BERT Bidirectional Encoder Representations from Transformers. 2, 4, 5, 10

CNN Convolutional Neural Network. 5, 7

ERASER Evaluating Rationales And Simple English Reasoning. 6, 8, 15

GloVe Global Vectors for Word Representation. 6

HASOC Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. 1–3, 10, 47

HITL human-in-the-loop. 10, 11

LIME Local Interpretable Model-Agnostic Explanations. 6, 8

LSTM Long Short-Term Memory. 5

NLP Natural Language Processing. 1, 7, 9

POTATO exPlainable infOrmation exTrAcTion framewOrk. 2, 4, 10, 11, 16, 23

RNN Recurrent Neural Network. 7

RQ research question. 13, 44

SHAP SHapley Additive exPlanations. 8

SVM Support Vector Machine. 2, 5

UD Universal Dependencies. 10, 23

URL Uniform Resource Locator. 13

Bibliography

[ACLM21] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection, 2021. `arXiv:2103.11800`.

[ADY^{+}21] Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5868–5876, 2021.

[BBC^{+}13] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, 2013.

[BBF^{+}19] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019.

[BDP17] Tiberiu Boros, Stefan Daniel Dumitrescu, and Sonia Pipa. Fast and accurate decision trees for natural language processing tasks. In *RANLP*, pages 103–110, 2017.

[Cas16] Davide Castelvecchi. Can we open the black box of AI? *Nature News*, 538(7623):20, 2016.

[Chu17] Kenneth Ward Church. Word2Vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[Cla19] Clark, Christopher and Lee, Kenton and Chang, Ming-Wei, and Kwiatkowski, Tom and Collins, Michael, and Toutanova, Kristina. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

[CRLB18] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc., 2018. URL: `http://papers.nips.cc/paper/8163-e-snli-natural-language-inference-with-natural-language-explanations.pdf`.

[CRT20] Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and characterizing human rationales. *arXiv preprint arXiv:2010.04736*, 2020.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DJR$^+$] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. https://www.eraserbenchmark.com/.

[DJR$^+$19] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429*, 2019.

[DMMNZ21] Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. Universal dependencies. *Computational linguistics*, 47(2):255–308, 2021.

[DQA$^+$20] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[ger19] Germeval Task 2, 2019 — Shared Task on the Identification of Offensive Language . Online, 2019. Last accessed: 2022-10-05. URL: `https://projects.fzai.h-da.de/iggsa/germeval/`.

[GF17] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[GKRR21] Kinga Gémes, Ádám Kovács, Markus Reichel, and Gábor Recski. Offensive text detection on english twitter with deep learning models and rule-based systems. In *Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org*, 2021.

[Goo20] Michael Wayne Goodman. Penman: An open-source library and tool for amr graphs. In *Proceedings of the 58th Annual Meeting of the Association* *for Computational Linguistics: System Demonstrations*, pages 312–319, 2020.

[has19] HASOC 2019. Hate Speech and Offensive Content Identification in Indo-European Languages. Online, 2019. Last accessed: 2022-10-05. URL: `https://hasocfire.github.io/hasoc/2019/index.html`.

[has20] HASOC 2020. Hate Speech and Offensive Content Identification in Indo-European Languages. Online, 2020. Last accessed: 2022-10-05. URL: `https://hasocfire.github.io/hasoc/2020/index.html`.

[has21] HASOC 2021. Hate Speech and Offensive Content Identification in Indo-European Languages. Online, 2021. Last accessed: 2022-10-05. URL: `https://hasocfire.github.io/hasoc/2021/index.html`.

[hat19] SemEval 2019 Task 5 - Shared Task on Multilingual Detection of Hate. Online, 2019. Last accessed: 2022-10-05. URL: `https://competitions.codalab.org/competitions/19935`.

[JG20] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

[KAM+15] András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado, June 2015. Association for Computational Linguistics.

[KGIR22] Ádám Kovács, Kinga Gémes, Eszter Iklódi, and Gábor Recski. POTATO: exPlainable infOrmation exTrAcTion framewOrk. *arXiv preprint arXiv:2201.13230*, 2022.

[Kov22] Ádám Kovács. Using POTATO for interpretable information extraction. Online, 2022. Last accessed: 2022-10-05. URL: `https://medium.com/towards-data-science/using-potato-for-interpretable-information-extraction-f2081a717eb7`.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[LWLQ21] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *arXiv preprint arXiv:2106.04554*, 2021.

[MMKMC20] Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, pages 29–32, 2020.

[MMS+21] Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*, 2021.

[MPH+19] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, 2019.

[MSY+21] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

[RG19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL: `https://arxiv.org/abs/1908.10084`.

[RMXS19] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL2019)*, 2019. URL: `https://arxiv.org/abs/1906.02361`.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[SS19] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[SSR+19] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. Overview of germeval task 2, 2019 shared task on the identification of offensive language. 2019.

[Sut16] Shan Suthaharan. Support vector machine. In *Machine learning models and algorithms for big data classification*, pages 207–235. Springer, 2016.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WBM18] Bernhard Waltl, Georg Bonczek, and Florian Matthes. Rule-based information extraction: Advantages, limitations, and perspectives. *Jusletter IT (02 2018)*, 2018.

[XMW+18] Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. Emo2vec: Learning generalized emotion representation by multi-task training. *arXiv preprint arXiv:1809.04505*, 2018.

[ZSW22] Mengdi Zhang, Jun Sun, and Jingyi Wang. Which neural network makes more explainable decisions? An approach towards measuring explainability. *Automated Software Engineering*, 29(2):1–26, 2022.

[ZWM19] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations, 2019. `arXiv:1909.10681`.