DELFT UNIVERSITY OF TECHNOLOGY

MASTERS THESIS

Value-Sensitive Rejection of Machine Learning Predictions for Hate Speech Detection

Author:
Philippe Lammerts

Thesis advisor:
Prof. dr. ir. G.J.P.M. Houben
Delft University of Technology

Daily supervisor:
Dr. ir. J. Yang
Delft University of Technology

Co-daily supervisors:
Dr. ir. Y-C. Hsu
University of Amsterdam

Ir. P. Lippmann
Delft University of Technology

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Web Information Systems Group - Crowd Computing
Software Technology

October 13, 2022

DELFT UNIVERSITY OF TECHNOLOGY

Abstract

Electrical Engineering, Mathematics and Computer Science
Software Technology

Master of Science

Value-Sensitive Rejection of Machine Learning Predictions for Hate Speech Detection

by Philippe Lammerts

Hate speech detection on social media platforms remains a challenging task. Manual moderation by humans is the most reliable but infeasible, and machine learning models for detecting hate speech are scalable but unreliable as they often perform poorly on unseen data. Therefore, human-AI collaborative systems, in which we combine the strengths of humans' reliability and the scalability of machine learning, offer great potential for detecting hate speech. While methods for task handover in human-AI collaboration exist that consider the costs of incorrect predictions, insufficient attention has been paid to estimating these costs. In this work, we propose a value-sensitive rejector that automatically rejects machine learning predictions when the prediction's confidence is too low by taking into account the users' perception regarding different types of machine learning predictions. We conducted a crowdsourced survey study with 160 participants to evaluate their perception of correct, incorrect and rejected predictions in the context of hate speech detection. We introduce magnitude estimation, an unbounded scale, as the preferred method for measuring user perception of machine predictions. The results show that we can use magnitude estimation reliably for measuring the users' perception. We integrate the user-perceived values into the value-sensitive rejector and apply the rejector to several state-of-the-art hate speech detection models. The results show that the value-sensitive rejector can help us to determine when to accept or reject predictions to achieve optimal model value. Furthermore, the results show that the best model can be different when optimizing model value compared to optimizing more widely used metrics, such as accuracy.

Acknowledgements

This work would not have been possible without the help of several people. First and foremost, I am extremely grateful to my supervisors, dr. Jie Yang, dr. Yen-Chia Hsu, and Philip Lippmann, for their excellent guidance and support throughout the project, contributions to the paper, and valuable knowledge and feedback. Special thanks to the Delft University of Technology for their resources and financial support for conducting the experiments. I would also like to thank my thesis committee and everyone who participated in the survey study. Finally, I would like to express my most profound appreciation to my family and friends, in particular Daan, Henk-Jan, Martijn, Michelle, Nick, and Simon, for their support.

Contents

List of Abbreviations

ANOVA	**an**alysis **o**f **va**riance
BERT	**b**idirectional **e**ncoder **r**epresentations from **t**ransformers
BOW	**b**ag-**o**f-**w**ords
CNN	**c**onvolutional **n**eural **n**etwork
DL	**d**eep **l**earning
ECE	**e**xpected **c**alibration **e**rror
FN	**f**alse **n**egative
FP	**f**alse **p**ositive
KDE	**k**ernel **d**ensity **e**stimation
LDA	**l**atent **d**irichlet **a**llocation
LR	**l**ogistic **r**egression
LSTM	**l**ong **s**hort-**t**erm **m**emory
ML	**m**achine **l**earning
MOE	**m**argin **o**f **e**rror
NLP	**n**atural **l**anguage **p**rocessing
PBT	**p**opulation **b**ased **t**raining
PDF	**p**robability **d**ensity **f**unction
POS	**p**art-**o**f-**s**peech
RR	**r**ejection **r**ate
SVM	**s**upport **v**ector **m**achine
TF-IDF	**t**erm **f**requency-**i**nverse **d**ocument **f**requency
TN	**t**rue **n**egative
TP	**t**rue **p**ositive
VSD	**v**alue-**s**ensitive **d**esign

Chapter 1

Introduction

The amount of hateful content spread online on social media remains a significant problem. Ignoring its presence can harm people and even result in actual violence and other conflicts (Balayn et al., 2021; Council of Europe, n.d.). Many news articles exist about events where hate spread on online platforms leads to acts of violence (Ingram, 2018; Mashal et al., 2022; Mozur, 2018; Müller & Schwarz, 2021). One research paper found a connection between hateful content on Facebook containing anti-refugee sentiment and hate crimes against refugees by analyzing social media usage in multiple municipalities in Germany (Müller & Schwarz, 2021). Governmental institutions and social media companies are becoming more aware of these risks and are trying to combat hate speech. For example, the European Union developed a Code of Conduct for countering illegal hate speech in cooperation with large social media companies such as Facebook and Twitter (European Commission, 2016). This Code of Conduct requests companies to prohibit hate speech and report yearly progress (European Commission, 2016). The most recent report from 2021 stated that Twitter only removed 49.5% of all hateful content on its platform. Facebook is most successful in removing hate speech, claiming to have removed 70.2% of all hateful content in 2021 (European Commission, 2016). However, one article found in internal communication from Facebook that this percentage is much lower, around 3-5% (Giansiracusa, 2021). Therefore, hate speech detection remains a complex problem that even large institutions have not solved yet.

1.1 Problem statement

Currently, people rely on reactive and proactive content moderation methods to detect hate speech (Klonick, 2018). Reactive moderation is when social media users flag hateful content (also known as reporting) (Klonick, 2018). Proactive moderation is done automatically using detection algorithms or manually by a group of human moderators (Klonick, 2018). There exist different methods for automatically detecting hateful content. Most use machine learning (ML) algorithms since these tend to be the most promising for their detection performance at a large scale (Balayn et al., 2021; Fortuna & Nunes, 2018). These algorithms range from traditional ML methods, such as support vector machines (SVM) or decision trees, to deep learning (DL) algorithms (Fortuna & Nunes, 2018).

However, both proactive and reactive moderation methods have their limitations. Proactive manual moderation of hateful content is still the most reliable solution, but it is simply infeasible because of the large amount of content generated by the many users (Balayn et al., 2021). Reactive moderation solves this problem since the users can report hate speech themselves. Although, the problem stays that users are still exposed to hateful content for some time. Proactive automatic moderation using automated detection algorithms allows large amounts of data to be

checked quickly without the involvement of humans. However, these algorithms are unreliable as they often perform poorly on deployment data (Arango et al., 2019; Balayn et al., 2021; Gröndahl et al., 2018). One study found that the F1 scores reduce significantly (69% F1 score drop in the worst case) when training a hate speech detection model on one dataset and evaluating it using another (Gröndahl et al., 2018). Furthermore, one paper found that most research in hate speech detection overestimates the performance of the automated detection methods (Arango et al., 2019). Likewise, the authors noticed significant performance drops when the detection algorithms were trained on one dataset and evaluated on another (Arango et al., 2019).

Therefore, humans and machines should work together to detect hate speech. ML models should detect hateful content automatically, and humans should make the final decisions (*human-in-the-loop*) when the model is not confident enough (Hendrickx et al., 2021; Woo, 2020). The challenge is to determine when we can accept ML predictions and when we need to reject them and defer them to a human moderator. We focus on deciding whether to accept or reject ML predictions by taking the context-dependent *values* into account. Several papers advocate for integrating context-dependent values into the design of human-AI systems (Casati et al., 2021; Cummings, 2006; Sayin et al., 2021; Umbrello & Van de Poel, 2021; Zhu et al., 2018). There are benefits of correct predictions (positive value), costs of incorrect predictions (negative value), and costs of rejecting predictions. More specifically, the values for false negative (FN) predictions, labelling something as non-hateful when it is, and false positive (FP) predictions, labelling something as hateful when it is not, might differ. We should weigh these values according to the task of hate speech detection and incorporate them in the decision of accepting or rejecting ML predictions (Sayin et al., 2021). However, value is an abstract term and can be interpreted from different perspectives, such as economic or social, by different stakeholders, such as the social media companies or the social media users (Cummings, 2006; Umbrello & Van de Poel, 2021; Zhu et al., 2018). In this project, we mainly focus on integrating human-centred social values from the perspective of social media users since they are the most affected by the consequences of hate speech. We also focus on hate speech detection as a binary classification problem, where a prediction for a social media post is either positive (hateful) or negative (neutral).

1.2 Research questions

The idea of most ML models with a reject option is that we reject predictions when the model's confidence is too low. First, we need a metric that measures the total value of an ML model with a reject option based on the context-dependent values. By optimizing the value of this metric, we can retrieve an optimal confidence threshold that we can use to determine when we can accept ML predictions or if we need to reject them. Second, we need to find out how we can retrieve context-dependent values for the task of hate speech detection. The goal is to retrieve the value ratios by which we mean, for example, the ratio between an FP and an FN prediction. Therefore, our research questions are as follows:

RQ How can we reject predictions of machine learning models in a value-sensitive manner for hate speech detection?

- **SRQ1** How can we measure the total value of machine learning models with a reject option?

- **SRQ2** How can we determine the value ratios between rejections and true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions?

1.3 Our work

This thesis research tackled the problems of proactive moderation by creating a human-AI solution where the advantages of humans (cognitive abilities and ability to make judgements) and machines (automation and performance) are combined (Woo, 2020). We did this by proposing a *value-sensitive* rejector for detecting hate speech that rejects ML predictions when the prediction's confidence is too low based on human-centred values. To the best of our knowledge, ML with rejection has not been used in hate speech detection before.

In this thesis, we conducted a survey study where we recruited 160 participants to evaluate their perception of 40 different hate speech detection scenarios. Each scenario simulates either a TP, TN, FP, FN, or rejected prediction in the context of hate speech detection. We carefully selected 40 social media posts through an extensive content analysis procedure for creating the scenarios. We proposed using the magnitude estimation (ME) rating scale for retrieving the value ratios by measuring the user perception of all scenarios. We validated the ME scale by conducting a separate survey study with a bounded scale comprising 100 rating levels, called the 100-level scale. The results show a high inter-rater agreement between the participants of the ME survey, indicating that ME is suitable for retrieving human-centred values. We also found that users tend to agree more with correct predictions than the degree of disagreement with incorrect predictions, implying that social media users highly appreciate correct predictions made by the social media platform. Additionally, we found that participants agree more with each other for incorrect predictions than for correct predictions, indicating a strong consensus over the harm. Finally, analysis of the demographical features showed that for most scenarios, there are no differences in the user perception between different demographic groups.

We created a value-sensitive metric that measures the total value of an ML model for some confidence threshold based on the value ratios and a set of predictions. We can convert any ML model into a value-sensitive rejector by finding the optimal confidence rejection threshold for which the value-sensitive metric achieves the maximum value. We applied the value-sensitive rejector, with the value ratios from the survey study as its input, to two different datasets and three state-of-the-art hate speech classification models: a traditional, a deep learning, and a transformer model. We denote the first dataset as the *seen* dataset, a test dataset from the same source as the training dataset of the models. We denote the second as the *unseen* dataset, a test dataset from a different source, to simulate how the models would perform in the real world on new data. The results demonstrate that the value-sensitive rejector can be beneficial for hate speech detection since we show that we maximize the values of several hate speech detection models by rejecting predictions. The results with the *unseen* data show that hate speech detection models are susceptible to bias, which confirms the findings from related studies. Finally, the results show that when selecting the optimal model, using the value of our value-sensitive metric as the optimization target might return different results than using accuracy.

1.4 Contributions

In summary, we make the following contributions:

- We introduce the concept of rejecting machine learning predictions into the task of hate speech detection;
- We introduce the magnitude estimation scale for measuring user perception to correct and incorrect machine learning predictions;
- We present a value-sensitive metric for measuring the total value of a machine learning model with a reject option;
- We demonstrate through a survey study that the magnitude estimation scale is suitable for retrieving the value ratios of TP, TN, FP, FN, and rejected predictions in the context of hate speech detection;
- We demonstrate that our value-sensitive rejector can guide us in determining when to accept or reject machine learning predictions to obtain optimal model values;

1.5 Thesis outline

In this thesis report, we first discuss the related work in chapter 2. Then in chapter 3, we present the design of the value-sensitive rejector. Chapter 4 explains the design of the survey study. In chapter 5, we present the results of the experiments of the survey study and the value-sensitive rejector. Finally, chapter 6 discusses the results, and chapter 7 contains the conclusion.

Chapter 2

Related work

In this chapter, we first define hate speech in section 2.1 and explain why it is such a challenging topic to tackle, especially from a computer science perspective. Then, we give an overview of the state-of-the-art solutions for automatic hate speech detection in section 2.2. In section 2.3, we discuss the different methods of ML with rejection. Section 2.4 discusses the shortcomings of standard machine metrics, such as accuracy, to evaluate detection systems and why human-centred metrics such as ours are promising. Finally, we discuss the main challenges of assessing the values of (in)correct and rejected predictions in the hate speech domain.

2.1 Hate speech: definition and challenges

Different types of online conflictual languages exist, such as cyberbullying, offensive language, toxic language, or hate speech, and come with varying definitions from domains such as psychology, political science, or computer science (Balayn et al., 2021). We can broadly define *hate speech* as "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" (Balayn et al., 2021; Davidson et al., 2017). It differs from other conflictual languages since it focuses on specific target groups or individuals (Balayn et al., 2021).

Balayn et al. (2021) identified the mismatch between the formalization of hate speech and how people perceive it. Many factors influence how people perceive hate speech, such as the content itself and the characteristics of the target group and the observing individual, such as sex, cultural background, or age (Balayn et al., 2021). We can identify this mismatch in other related work from which there appears to be low agreement among humans regarding annotating hate speech (Fortuna & Nunes, 2018; Ross et al., 2017; Waseem, 2016). Ross et al. (2017) reported low inter-rater reliability scores (Krippendorff's alpha values of around $0.2 - 0.3$) in a study where they asked humans about the hatefulness and offensiveness of a selection of tweets. They also found that the inter-rater reliability value does not increase when showing a definition of hate speech to the human annotators beforehand. Waseem (2016) found a slight increase in the inter-rater reliability when considering annotations of human experts only, but it remained low overall.

In the hate speech domain, we must be careful with creating biased detection systems trained on biased datasets. Hate speech datasets such as Waseem and Hovy (2016) or Basile et al. (2019) collected their data using specific keywords that can introduce *sample retrieval* bias and annotated their data using only three independent annotators, which might result in *sample annotation* bias (Balayn et al., 2021). Automated classification models will likely become biased in their predictions if we train them on biased datasets (the *garbage in, garbage out* principle). This phenomenon becomes most notable when applying pre-trained classification models to new and

unseen data. For example, Gröndahl et al. (2018) and Arango et al. (2019) report significant drops in F1 scores when training a hate speech classification model on one dataset and evaluating it on another. Gröndahl et al. (2018) found that the F1 score reduces by 69% in the worst case and that the model choice does not affect the classification performance as much as the dataset choice. Arango et al. (2019) replicated several state-of-the-art hate speech classification models and found that most studies overestimate the classification performance. These results further strengthen our stance that we should not detect hate speech solely by machines but rather by a human-in-the-loop approach.

2.2 Automatic hate speech detection

This section will list the literature's state-of-the-art natural language processing (NLP) techniques for automatic hate speech detection. This project focuses on hate speech detection as a binary text classification problem. The goal is to label texts from social media platforms as either hateful or not hateful. Several excellent surveys outlined the different detection methods (Fortuna & Nunes, 2018; Schmidt & Wiegand, 2019). First, we will discuss the different features used in the classification models. Then, we will state the most used classification models ranging from supervised to unsupervised learning.

Commonly used features are bag-of-words (BOW) (Greevy & Smeaton, 2004), character/word N-grams (Waseem & Hovy, 2016), lexicon features (Xiang et al., 2012), term frequency-inverse document frequency (TF-IDF) (Badjatiya et al., 2017; Davidson et al., 2017; Rodriguez et al., 2019), part-of-speech (POS) (Greevy & Smeaton, 2004), sentiment analysis (Rodriguez et al., 2019), topic modelling (e.g. latent dirichlet allocation (LDA)) (Xiang et al., 2012), meta-information (e.g. location) (Waseem & Hovy, 2016), or word embeddings (Agrawal & Awekar, 2018; Badjatiya et al., 2017). Greevy and Smeaton (2004) found that the classification performance is higher with BOW features than with POS features. Waseem and Hovy (2016) found that character N-gram achieves higher classification performance than word N-gram. They also found that using demographic information such as the location does not improve the results significantly. Xiang et al. (2012) used a lexicon feature (whether a social media post contains an offensive word or not) and the topic distributions from an LDA analysis. Rodriguez et al. (2019) used TF-IDF and sentiment analysis to detect and cluster topics on Facebook pages that are likely to promote hate speech. Badjatiya et al. (2017) experimented with different word embeddings: fastText[1], GloVe[2], and random word embeddings. They found that using pre-trained word embeddings such as GloVe does not result in better classification performance than using random embeddings.

Most studies use supervised learning techniques that range from traditional ML to deep learning (DL) classification models, and a few use unsupervised learning techniques to cluster social media posts. Support vector machine (SVM) (Davidson et al., 2017; Greevy & Smeaton, 2004; Xiang et al., 2012) and logistic regression (LR) (Davidson et al., 2017; Waseem & Hovy, 2016) are the most popular traditional ML techniques for hate speech detection. Davidson et al. (2017) found that SVM and LR perform significantly better than other traditional ML techniques, such as naive Bayes, decision trees, and random forests. Badjatiya et al. (2017) experimented with various configurations of word embeddings and two DL models: a convolutional

[1] https://fasttext.cc/
[2] https://nlp.stanford.edu/projects/glove/

neural network (CNN) and a long short-term memory (LSTM) model. They found that CNN performs better than LSTM. Given the recent popularity of bidirectional encoder representations from transformers (BERT) models (Devlin et al., 2018) in the NLP field, studies such as Alatawi et al. (2021) found that BERT models achieve slightly better classification performance than DL models. Rodriguez et al. (2019) use the unsupervised learning method, k-means clustering, to cluster social media posts to identify topics that potentially promote hate speech. Based on the findings of these studies, we will experiment with three models in our project: LR, CNN, and DistilBERT (a lightweight version of BERT (Sanh et al., 2019)).

2.3 Machine learning with rejection

Several related studies promoted the concept of rejecting ML predictions when the risk of producing an incorrect prediction is too high so that a human gives the final judgement instead (Hendrickx et al., 2021; Sayin et al., 2021; Woo, 2020). Hendrickx et al. (2021) identified three ways of rejecting ML predictions: *separated*, *integrated*, and *dependent*. A separated rejector decides beforehand whether a data sample needs to be handled by the classification model or not (Hendrickx et al., 2021). An integrated rejector forms one whole with a classification model that we often train simultaneously (Hendrickx et al., 2021). A dependent rejector analyzes the output of the classification model to determine whether to reject a prediction or not (Hendrickx et al., 2021). Several studies have applied the reject option using one of the abovementioned architectures (Coenen et al., 2020; De Stefano et al., 2000; Geifman & El-Yaniv, 2017, 2019; Grandvalet et al., 2008).

Coenen et al. (2020) developed a *separated* rejector that rejects data samples before passing them to the classification model. They used different outlier detection techniques, such as the one-class SVM, to detect data samples unfamiliar with the training data (Coenen et al., 2020).

Dependent rejectors are the most commonly used (De Stefano et al., 2000; Geifman & El-Yaniv, 2017; Grandvalet et al., 2008). Grandvalet et al. (2008) experimented with SVMs with a reject option. Geifman and El-Yaniv (2017) developed a dependent rejector that rejects data samples based on a predefined maximum risk value and the coverage accuracy of the classification model. De Stefano et al. (2000) were among the first to develop a dependent rejector for neural networks. The authors developed a confidence metric for determining the optimal rejection threshold (De Stefano et al., 2000). This threshold is calculated based on a set of predictions with their corresponding confidence values and a set of cost values: the cost of incorrect, correct, and rejected predictions (De Stefano et al., 2000).

Geifman and El-Yaniv (2019) developed an *integrated* rejector by extending the work from Geifman and El-Yaniv (2017). They integrated the reject option in the training phase of a DL classification model by including a selection function in the last layer of the DL model.

In this work, we apply the dependent way since it allows for using the reject option in any classification model (Hendrickx et al., 2021). As opposed to the integrated way, by following the dependent way, we are free to use any classification model, and we do not have to retrain the underlying model whenever we make modifications to the dependent rejector. We believe that the separated way is not optimal either since we still want to decide whether to accept or reject predictions based on the output of the classification model. The most relevant work in dependent rejectors is from De Stefano et al. (2000) since their confidence metric considers the

value of (in)correct and rejected predictions. While their metric measures only the effectiveness of the reject option and is based on the values of correct, incorrect, and rejected predictions, our metric measures the total value of the ML model with the reject option and is based on the values of TP, TN, FP, FN, and rejected predictions. While they experimented with a range of different cost values, we go further by employing an emperical approach, which determines the cost values based on how users feel regarding machine predictions using a survey study with crowd workers. Therefore, we obtain a rejection threshold that captures the implications of machine predictions from a human perspective.

2.4 Evaluation metrics

Most hate speech-related studies evaluate their classification methods using standard *machine* metrics such as accuracy, precision, recall, or F1. Classification models with a reject option are often evaluated by analyzing the model's accuracy and coverage. Nadeem et al. (2009) proposed using accuracy-rejection curves to plot the trade-off between accuracy and coverage so that different classification models with a reject option can be compared. Casati et al. (2021), Olteanu et al. (2017), and Röttger et al. (2020) recognized the shortcomings of machine metrics, such as accuracy and found a gap in the evaluation of hate speech detection systems.

Röttger et al. (2020) found it hard to identify the weak points of classification models using machine metrics, such as accuracy. Therefore, the authors presented a suite that consists of 29 carefully selected functional tests to help identify the model's weaknesses (Röttger et al., 2020). Each test checks criteria, such as coping with spelling variations or detecting neutral content containing slurs (Röttger et al., 2020). Our approach is different since we focus on measuring the value of classification models with a reject option.

Olteanu et al. (2017) promote using *human-centred* metrics that measure the human-perceived value of hate speech classification models. They found that for the same precision values, the perceived value changes depending on the user characteristics and the type of classification errors (an offensive tweet labelled as hate (low impact) and a neutral tweet labelled as hate (high impact)) (Olteanu et al., 2017).

Casati et al. (2021) propose to develop new metrics for evaluating ML models with a reject option that considers domain-specific values. Our work aligns with the latter two studies since we create a human-centred metric for evaluating hate speech classification models with a reject option that incorporates human value derived from a survey study.

2.5 Value assessment

Fjeld et al. (2020) outlined eight principles of AI systems, such as *fairness and discrimination* (e.g. preventing algorithmic bias), *human control of technology* (e.g. the system should request help from the human user in difficult situations), and *promotion of human values* (e.g. we should integrate human value in the system). Sayin et al. (2021) and Casati et al. (2021) suggest we should identify context-specific *values* and incorporate them in the design of a hybrid human-AI system. We adhere to the suggestions of these studies in our project since we develop a hate speech classification model with a reject option that incorporates human value.

As explained in the Introduction, we have costs of incorrect and rejected predictions and gains of correct predictions. We can express the costs of incorrect (FP and

FN) and rejected predictions as negative values and the gains of correct (TP and TN) predictions as positive values. We should weigh these values according to the task of case hate speech detection (Sayin et al., 2021). However, value is a broad term, and its definition depends heavily on the context.

Several works discuss the value-sensitive design (VSD) approach that describes how different types of value, such as privacy, can be integrated into a socio-technical system's design (Cummings, 2006; Umbrello & Van de Poel, 2021; Zhu et al., 2018). According to the VSD approach, it is critical to understand the system's stakeholders, and we can retrieve their values either *conceptually* (e.g. from literature) or *empirically* (e.g. through survey studies) (Cummings, 2006; Umbrello & Van de Poel, 2021; Zhu et al., 2018).

We consider two different stakeholders: the social media platforms and the users. The goal is to find out whether we can retrieve the value ratios between rejection, FP, FN, TP, and TN predictions from the perspective of both stakeholders. We would like to know whether an FN prediction is, for example, two times worse than an FP prediction. The main challenge is to express all values using a single unit. First, we could define the values using a quantitative measure, such as time or money spent/saved. Second, we could define the values using a qualitative measure, for example, by analyzing people's stance towards the consequence of incorrect predictions in hate speech detection.

In this section, we try to assess the values of both stakeholders empirically and conceptually and explain why we eventually go for an empirical analysis of the values of social media users only.

2.5.1 Quantitative assessment

In this section, we explain the difficulties of using quantitative measurements to define the values of TP, TN, FP, FN, and rejected predictions in hate speech detection. We do this by following the conceptual approach for both stakeholders by looking at some related work to see if the empirical approach is possible.

First, we look at the social media company as a stakeholder. We can retrieve the value of rejection by looking at how much time a human moderator spends on average to check whether some social media post contains hateful content or not. We can convert this into money by considering the moderator's salary. We could also argue that the value of a TP and a TN prediction is equal to the negative value of rejection since we saved human effort by having the classification model produce a correct prediction. The problem, however, starts to arise when we look at the FP and the FN predictions. How can we express the values of FP and FN predictions regarding money or time saved/spent? The main problem is that most social media companies are not transparent about moderating hate speech (Klonick, 2018). So it is infeasible to assess the values of social media companies either conceptually or empirically. When looking at the consequences of FN predictions, we can also look at governmental fines. For example, Germany approved a plan where social media companies can be fined up to 50 million euros if they do not remove hate speech in time ("Social media firms faces huge hate speech fines in Germany", 2017). However, this is location-specific, and it is unclear how this applies to individual cases of hate speech. Defining the value of FP predictions is even more difficult. It is unclear how filtering out too much content would affect the company regarding money/time lost. Therefore, we abstain from estimating the values where the companies are the main stakeholders.

Second, we look at the social media users as a stakeholder. Both FP and FN predictions have negative consequences on the users. Having too many FP predictions might violate the value of Freedom of Speech since we are filtering out non-hateful posts and, therefore, we cause suppression of free speech. One paper found through a survey that most people think some form of hate speech moderation is needed, but they also worry about the violation of freedom of speech (Olteanu et al., 2017). Having too many FN predictions might harm individuals or even result in acts of violence (Council of Europe, n.d.). Therefore, we must figure out how to weigh the values of FP and FN predictions accordingly. We abstain from using time as a unit since it does not make sense to express the consequences of hate speech or the benefits of freedom of speech in time. Therefore, we want to look at the value of freedom of speech and hate speech from an economic perspective. However, we noticed a lack of research in this area. There is one paper where they tried to develop an economic model for free political speech by looking at the First Amendment to the United States Constitution (Posner, 1986). The First Amendment restricts the government from creating laws that could, for example, violate Freedom of Speech ("The Constitution", n.d.). Posner (1986) explained that the lack of research in this area is because most economists do not dive into the legal domain regarding free speech, and free speech legal specialists refrain from doing economic analysis (Posner, 1986). The proposed economic model from the paper includes the cost of harm and the probability that speech results in violence (Posner, 1986). However, the authors do not elaborate on how we can define the probability and the costs. Another paper did speculate on this topic by explaining why doing a cost-benefit analysis of free speech is almost impossible (Sunstein, 2019). The authors explained that there are too many uncertainties (Sunstein, 2019). We can assume that there are values of free speech, but it is too difficult to quantify them (Sunstein, 2019). Terrorist organizations use free speech to recruit people and call for acts of violence online (Sunstein, 2019). At the same time, most other hateful posts will never result in actual acts of violence (Sunstein, 2019). Therefore, value assessment using quantitative measurements is already tricky for specific cases, let alone in general. There is a nonquantifiable risk that acts of violence will happen in the unknown future (Sunstein, 2019). However, suppose we know this probability, there are still too many uncertainties. To calculate the actual costs of hate speech (the FN predictions), we also need to know the number of lives at risk and how we should quantify the value of each life (Sunstein, 2019). The authors claim that analyzing the benefits of free speech is even more challenging (Sunstein, 2019). They conclude their work by saying that there are too many problems to empirically evaluate the costs and benefits of hate speech detection (Sunstein, 2019).

Therefore, we believe that using quantitative measurements, such as money, is impossible to assess the values of predictions for both stakeholders in hate speech detection.

2.5.2 Qualitative assessment

From section 2.5.1, we concluded that from related work, it appears that we cannot retrieve the quantitative values conceptually and empirically. Instead, we will focus on the qualitative measurement of values: what is people's stance towards (in)correct and rejected predictions in hate speech detection? We only consider the social media users as the stakeholder in the qualitative assessment since they are the most affected by the consequences of hate speech detection. We will empirically

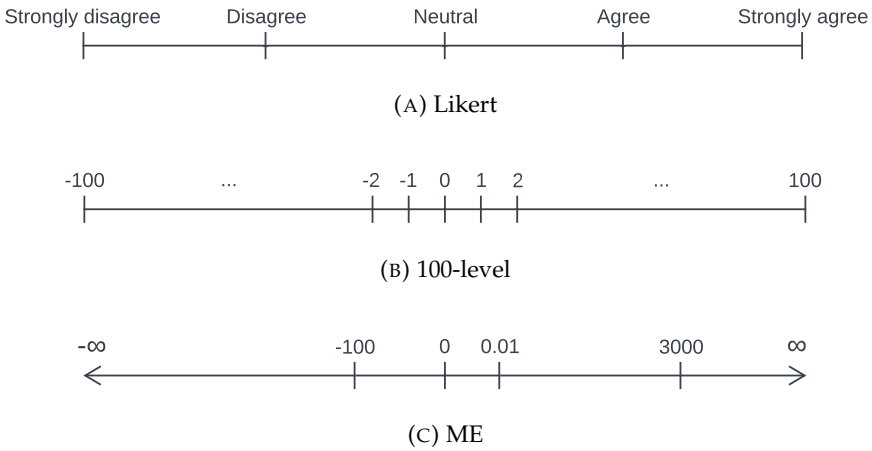
assess social media users' value through a survey. In our survey, we ask social media users what their stance (disagree-agree) is towards TP, TN, FP, FN, and rejected predictions in hate speech detection. Conceptual analysis is impossible since no related studies have tackled this problem. The closest work is from Ross et al. (2017), where the authors asked human subjects to rate a selection of tweets on hatefulness using a 6-point Likert scale and to indicate whether they think it should be banned from Twitter or not. Like Ross et al. (2017), we could use the Likert scale as our measurement scale. However, we first explain why Likert scales are unsuitable for retrieving ratio values. Then we explain why the magnitude estimation technique seems promising for our use case.

Likert

Likert scales are a common choice in academic research for retrieving the opinions of a group of subjects. Likert scales are multiple Likert-type questions (items) where subjects can answer questions with several response alternatives (Boone & Boone, 2012). For example, we could use a bipolar scale with seven response alternatives ranging from 'strongly disagree' to 'strongly agree', including a 'neutral' midpoint. Figure 2.1a shows an example of a five-point Likert item. However, there is much discussion in the literature about how we should analyze these Likert scales (Allen & Seaman, 2007; Boone & Boone, 2012; Murray, 2013; Norman, 2010). The scale of the questions is ordinal, which means that we know the responses' ranking, but we do not have an exact measurement of the distances between the response items (Allen & Seaman, 2007). For example, we know that 'strongly agree' is higher in rank than 'agree', but not the exact distance between the two responses and whether it is greater than the distance between the 'neutral' and the 'somewhat agree' responses. Therefore, we technically cannot use parametric statistics, such as calculating the mean, when analyzing the data (Allen & Seaman, 2007). Other papers argue that we can treat a Likert scale consisting of multiple Likert items as interval data; therefore, applying parametric statistics will not affect the conclusions (Boone & Boone, 2012; Murray, 2013; Norman, 2010). So, we can calculate mean scores for TP, TN, FP, FN, and rejected predictions and compare these with each other. For example, we can then verify that the mean value of FN predictions is smaller than the mean value of FP predictions and conclude that FN predictions are worse than FP predictions. Analyzing Likert scales would, at most, provide us with interval data (data for which we know the order, and we can measure the distances, but there is no actual zero point (Allen & Seaman, 2007)). However, we need to have ratio data in this project since we want to know the value ratios between the TP, TN, FP, FN, and rejected predictions.

Magnitude estimation

We concluded in the previous section that Likert scales are unsuitable since they do not provide ratio data. In this research, we want to experiment with the magnitude estimation (ME) technique. The ME technique originates from psychophysicists, where human subjects must give quantitative estimations of sensory magnitudes (Stevens, 1956). For example, in one experiment, human subjects are asked to assign any number that reflects their perception of the loudness of a range of sounds (Stevens, 1956). If the human subjects perceive the succeeding sound as twice as loud, they should assign a number to it that is twice as large. Researchers applied the ME technique to different types of physical stimuli (e.g. line length, brightness,



(A) Likert

(B) 100-level

(C) ME

FIGURE 2.1: Visualizations of all three rating scales: a five-point Likert item, the 100-level scale, and the ME scale.

or duration) and showed that the results are reproducible and that the data has ratio properties (Moskowitz, 1977). Other works have shown that the ME technique is also helpful for rating more abstract types of stimuli, such as judging the relevance of documents (Maddalena et al., 2017; Roitero et al., 2018), the linguistic acceptability of sentences (Bard et al., 1996), the strength of political opinions (Lodge et al., 1976; Lodge & Tursky, 1979), and the usability of system interfaces (McGee, 2004). Therefore, we think that ME is a promising method for retrieving the value ratios of the different types of predictions in hate speech detection.

The main advantage of ME is that it provides the ratio scale properties we need. Another advantage is that the scale is unbounded compared to other commonly used response scales, such as Likert. For example, suppose the subject provides a 'strongly disagree' judgment for the first stimulus. Suppose we then present an even worse stimulus. The subject is now limited to the response items in the Likert scale and can only give the same 'strongly disagree' judgement. We do not have this problem using ME because the subject is always free to assign a more significant value of disagreement. Figure 2.1c shows an example of a bipolar ME scale where any positive or negative numerical value is allowed, including decimal values. However, there are two drawbacks to using ME in our use case. First, we need to normalize the results since each subject uses a different range of values. Second, since ME has not been applied to the hate speech domain before, we need to validate the ME scale to verify that it measures what we want to know.

The data needs to be normalized since each subject can use any value they like. For example, one may give ratings using values of 1, 2, and 10, while another may use 100, 200, and 1000. Geometric averaging is the recommended approach for normalizing magnitude estimates since it preserves the ratio information (Maddalena et al., 2017; McGee, 2004; Moskowitz, 1977). However, as opposed to the unipolar scales (with only positive values) used by Bard et al. (1996) and McGee (2004) and Maddalena et al. (2017), we cannot apply geometric averaging to bipolar scales (disagree-agree). By including 0 (neutral) and negative values (disagree), we cannot use geometric averaging anymore because it uses log calculations (Moskowitz, 1977). Using the algorithmic mean is also not an option since it would destroy the ratio scale properties (Moskowitz, 1977). Therefore, we can normalize the magnitude estimates for bipolar scales by dividing all estimates of each subject by the maximum given value (Moskowitz, 1977). This way, all magnitude estimates are in the range

[-1, 1] while maintaining the ratio properties.

Most papers that use the ME method in a new domain apply some form of validation. Cross-modality validation is a technique that is often applied to validate the ME results (Bard et al., 1996). Psychophysicists compare the magnitude estimates to the physical stimuli by analyzing their correlation (Bard et al., 1996). In the case of estimating line lengths, we can easily vary the line length, for example, by showing a line that is twice as long as the previous line. Subjects can then estimate the line length using a number twice as large. However, this becomes more difficult in the social and psychology domains. In hate speech detection and other social science and psychology applications, we do not have an exact measure of the stimulus (Bard et al., 1996). However, related work has shown that ME is still a suitable technique for eliciting opinions about different types of non-physical stimuli (Bard et al., 1996; Lodge & Tursky, 1979; Maddalena et al., 2017; McGee, 2004). We can validate the magnitude estimates by adopting the cross-modality technique but instead compare judgements against judgements (Bard et al., 1996; Lodge & Tursky, 1979). Some papers analyze the correlation between different ME scales for validation, such as handgrip measurements or drawing lines (Bard et al., 1996; Lodge et al., 1976). Others compare ME with another validated scale that can be of any type. For example, in Maddalena et al. (2017), which is about judging the relevance of documents, the authors compared the ME scale with two validated ordinal scales for the same dataset (Maddalena et al., 2017). In Roitero et al. (2018), the authors applied cross-modality analysis between a bounded scale that consists of 100 levels (now known as the 100-level scale) and the ME scale and found that they were positively correlated. In our work, we follow the approach from Roitero et al. (2018), as we also validate our findings by checking the correlation between the ME scale and the 100-level scale. Figure 2.1b visualizes a bipolar 100-level scale.

Chapter 3

Value-sensitive rejector

As concluded in chapter 2, there is a need for *value-sensitive* metrics to measure ML models' performance, especially for social-technical applications such as hate speech detection. We also concluded that manual human moderation is the most effective and that most automatic hate speech detection methods do not perform well on unseen data. Therefore, in this project, we focus on creating a human-AI solution for detecting hate speech by rejecting ML predictions in a value-sensitive manner. We do this by taking the value ratios of TP, TN, FP, FN, and rejected predictions into account. Chapter 4 will explain how we assess these values. We assume that we know these values for the remaining part of this chapter.

In this chapter, we explain how we create a value-sensitive dependent rejector by introducing a value-sensitive confidence metric that measures the total value of an ML model with a reject option. In 3.1, we explain how we construct the confidence metric. In 3.2, we provide an overview of how we use the value-sensitive rejector, and in 3.3, we discuss how we apply the rejector to some state-of-the-art hate speech classification models. Refer to Appendix C for the source code of our value-sensitive rejector.

3.1 Value-sensitive metric

The idea of rejecting ML predictions using a confidence threshold is that for some threshold value τ in the range $[0, 1]$, we accept all predictions with confidence values greater than or equal to τ and reject all predictions with confidence values below τ. We use a confidence metric to find the optimal rejection threshold that is based on the work of De Stefano et al. (2000). Here, we introduce our confidence metric as the value function $V(\tau)$ that measures the total value of an ML model and rejection threshold τ. We can determine the optimal rejection threshold by finding the τ value for which $V(\tau)$ is the maximum. The value of $V(\tau)$ depends on the values of TP, TN, FP, FN, and rejected predictions, and we calculate it for a set of predictions with their corresponding confidence values and actual labels. We denote the values of TP, TN, FP, FN, and rejected predictions as V_{tp}, V_{tn}, V_{fp}, V_{fn}, and V_r, respectively. We derive the subsets of TP, TN, FP, and FN predictions from a set of predictions based on the predicted and actual labels.

We should be free to use any value for V_{tp}, V_{tn}, V_{fp}, V_{fn}, and V_r since we do not know which values will come from the survey study in chapter 4. However, for constructing our metric, we can define several conditions if we assume that V_{tp} and V_{tn} are positive values (gains) and V_{fp}, V_{fn}, and V_r are negative values (costs). For each τ value in $[0, 1]$, we would like to know whether the model with the reject option is more effective (increased $V(\tau)$) or less effective (decreased $V(\tau)$ value). We define the following conditions:

1. The value of incorrect predictions should be lower than that of rejected predictions. Otherwise, adopting the reject option serves no purpose.
2. Correct accepted predictions should increase the value of $V(\tau)$, while incorrect accepted predictions should decrease the value of $V(\tau)$.
3. Correct rejected predictions should decrease the value of $V(\tau)$, while incorrect rejected predictions should increase the value of $V(\tau)$.

We can formulate the first condition as follows:

$$\frac{V_{fp} + V_{fn}}{2} < V_r, \tag{3.1}$$

We can convert the latter two conditions into the following equations:

$$\frac{\partial V}{\partial F_{tp}} + \frac{\partial V}{\partial F_{tn}} > 0, \qquad \frac{\partial V}{\partial F^r_{tp}} + \frac{\partial V}{\partial F^r_{tn}} < 0, \tag{3.2a}$$

$$\frac{\partial V}{\partial F_{fp}} + \frac{\partial V}{\partial F_{fn}} < 0, \qquad \frac{\partial V}{\partial F^r_{fp}} + \frac{\partial V}{\partial F^r_{fn}} > 0, \tag{3.2b}$$

where F_p and F^r_p are the fractions of accepted and rejected predictions, respectively and $p \in [tp, tn, fp, fn]$. We create a linear $V(\tau)$ function and assume that the input values are known constants. Subsequently, we can formulate $V(\tau)$ as:

$$V(\tau) = \sum_p (V_p - V_r) F_p(\tau) + \sum_p (V_r - V_p) F^r_p(\tau), \tag{3.3}$$

where $p \in [tp, tn, fp, fn]$ and where $F_p(\tau)$ and $F^r_p(\tau)$ are the fractions of accepted and rejected predictions dependent on the rejection threshold τ. Conditions 3.2a are satisfied by default since we assume that V_{tp} and V_{tn} are positive and V_r is negative. Conditions 3.2b are satisfied since we assume that V_{fp}, V_{fn}, and V_r are negative and that condition 3.1 holds. We can retrieve the F_p and the F^r_p values by computing the integrals over the probability density functions (PDF) of the confidence values (denoted as x) of the predictions with type p. We compute the PDFs so that the calculation of the optimal rejection threshold is less sensitive to confidence outliers in the set of predictions. We denote F_p by taking the integral over the interval $[\tau, 1]$, and F^r_p by taking the integral over the interval $[0, \tau]$:

$$F_p(\tau) = \int_\tau^1 D_p(x)dx \qquad F^r_p(\tau) = \int_0^\tau D_p(x)dx, \tag{3.4}$$

where D_p is the PDF of all predictions of type p. By inserting the integrals from 3.4 into 3.3, we get our final value function:

$$V(\tau) = \sum_p (V_p - V_r) \int_\tau^1 D_p(x)dx + \sum_p (V_r - V_p) \int_0^\tau D_p(x)dx \tag{3.5}$$

We can now use 3.5 to calculate the total value of an ML model for all thresholds $\tau \in [0, 1]$. The theoretical optimal rejection threshold is equal to the τ value for which we achieve the maximum value of $V(\tau)$. We can find the optimal rejection threshold τ_O using the following formulation:

$$\tau_O \text{ where } V(\tau_O) = \max\{V(\tau) : \tau \in \mathbb{R} \wedge 0 \leq \tau \leq 1\} \tag{3.6}$$

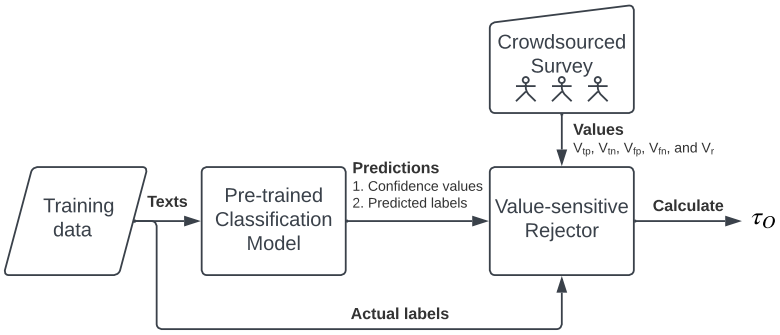


FIGURE 3.1: **Training phase:** flow diagram that visualizes how the value-sensitive rejector calculates the optimal rejection threshold τ_O.

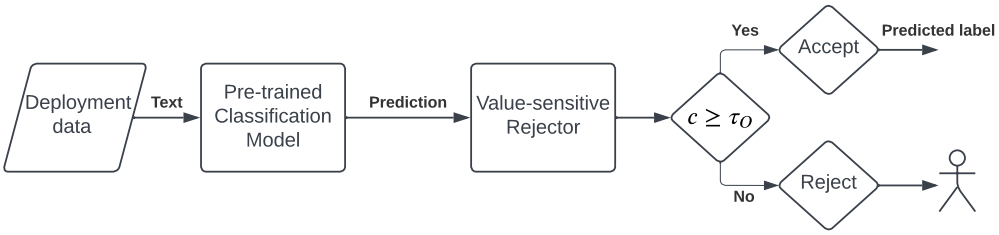


FIGURE 3.2: **Deployment phase:** flow diagram that visualizes how the value-sensitive rejector uses the optimal rejection threshold τ_O and the prediction confidence c to determine when to accept or reject a prediction from unseen data in deployment.

3.2 Overview of the value-sensitive rejector

This section provides an overview of how we use our value-sensitive rejector. We distinguish a training phase and a deployment phase of the rejector. In this project, we mainly focus on the training phase since we do not apply the rejector in the wild. Figures 3.1 and 3.2 visualize how we train the rejector and how we can use it in deployment to accept or reject predictions, respectively. In figure 3.1, we show the training phase of the rejector. In this phase, we use our value-sensitive metric from section 3.1 to calculate the optimal rejection threshold τ_O. We use the following inputs in this calculation: the values from the crowdsourced survey and a set of predictions that consist of the confidence values and the predicted and actual labels. Figure 3.2 shows how we can apply the trained rejector to unseen data in deployment. We accept all predictions for which the confidence value c is greater than or equal to the optimal rejection threshold τ_O and, otherwise, reject them so that a human moderator handles the prediction.

3.3 State-of-the-art

This section will explain how we apply the value-sensitive rejector to some of the state-of-the-art automatic hate speech detection models. In this experiment, we aim to find out three things. First, we want to determine how the value-sensitive rejector behaves on different models and datasets. Second, we want to know whether value-sensitive rejection can benefit hate speech detection. Finally, we compare the values

of our value-sensitive metric to the values of machine metrics such as accuracy and check whether they give different results.

3.3.1 Models

We experiment with three different hate speech detection models based on the findings from related work in section 2.2. The first model is a traditional ML model. We implement the LR model with character N-gram from Waseem and Hovy (2016) since this model achieved the best performance compared to other traditional ML models (Davidson et al., 2017). We select the second model, a DL model, based on the findings from Agrawal and Awekar (2018) and Badjatiya et al. (2017). We choose a CNN model initialized with random word embeddings since both studies found that this configuration provides state-of-the-art classification performance. We implement the CNN model based on the work of (Agrawal & Awekar, 2018). Finally, our third model is a BERT-based model, given its recent popularity in the NLP domain. We use the DistilBERT model since it is faster to train and smaller than BERT models while achieving similar performance (Sanh et al., 2019). We implement all models in Python. We implement the LR model with scikit-learn[1], the CNN model with TensorFlow[2], and the DistilBERT model with a combination of Hugging Face[3] and PyTorch[4]. We use Google Colab[5] to train all models.

3.3.2 Hyperparameter optimization

We perform hyperparameter optimization on all three models. For the CNN model, we apply random search to optimize the values of the learning rate, batch size, and the number of epochs. For the DistilBERT model, we apply population based training (PBT) (Jaderberg et al., 2017) implemented in Tune (Liaw et al., 2018), to optimize the values of the batch size, learning rate, and the number of epochs since Tune. PBT combines random search and hand tuning by discovering potentially optimal hyperparameter values along the way to reduce optimization time (Jaderberg et al., 2017). For the LR model, we use the LogisticRegressionCV model from scikit-learn that automatically optimizes the C value (inverse of regularization strength) of the LR model.

3.3.3 Calibration

The problem with most neural network models is that they are often not calibrated (Guo et al., 2017; Sayin et al., 2021). We define calibrated models as models where the confidence values of the predictions are equal to the probabilities that the predicted labels are correct. However, most neural networks tend to be sensitive to producing both low- and high-confident errors (Guo et al., 2017; Sayin et al., 2021). A well-calibrated model that achieves a low accuracy score can still be valuable since we can reject all low-confident incorrect predictions and only accept the high-confident correct predictions (Sayin et al., 2021). In our project, we aim to have calibrated models since calculating the optimal rejection threshold depends on the confidence values of the predictions.

[1] https://scikit-learn.org/
[2] https://www.tensorflow.org/
[3] https://huggingface.co/
[4] https://pytorch.org/
[5] https://colab.research.google.com/

Guo et al. (2017) experimented with different calibration methods. They evaluated the results using the expected calibration error (ECE), which measures the difference between the expected confidence and accuracy (Guo et al., 2017). They found that the temperature scaling method is the most effective. In temperature scaling, we divide the model's output logits with a temperature value of T to soften the probabilities of the final softmax function in the model's architecture (Guo et al., 2017). This T value is initially set to 1 and optimized by minimizing the negative log-likelihood (Guo et al., 2017). Please note that temperature scaling does not change the model's accuracy but only rescales the distribution of the confidence values (Guo et al., 2017).

As we experiment with two neural networks (DistilBERT and CNN), we apply temperature scaling to calibrate both models. However, calibration with temperature scaling does not guarantee perfect calibration. Therefore, high-confident incorrect predictions and low-confident correct predictions can still occur after calibration. Nevertheless, it is still valuable to calibrate the models since it also benefits human interpretation of the confidence values and, therefore, the interpretation of the optimal rejection threshold.

The Logistic Regression model is well-calibrated by default since, under the hood, it optimizes the log-loss function, which measures the difference between predicted confidence values and the actual labels. Therefore, we do not have to apply temperature scaling to the Logistic Regression model.

3.3.4 Datasets

We train all models on the Waseem and Hovy (2016) dataset consisting of 16K tweets labelled racist, sexist, or neutral. We converted the 'racist' and 'sexist' labels to 'hate' labels to create a binary classification setting. Furthermore, we split the dataset into a train and test dataset according to an 80:20 ratio. For the CNN and the DistilBERT models, we split the training set up into a training set and a validation set according to a 75:25 ratio. We use this validation set to calibrate the trained models by finding the optimal T value for the temperature scaling method. We preprocess the data by tokenizing all URLs, user mentions, and emojis since these do not contain any valuable information. We split all hashtags up into separate words using the WordSegment[6] library. The remaining parts of the preprocessing, such as removing whitespaces and stop words or the tokenization process, are dedicated to the different frameworks we use per model.

We apply the value-sensitive rejector to two test datasets: the *seen* and *unseen* dataset. The seen dataset is the test set from the Waseem and Hovy (2016) dataset. The unseen dataset is a test set from the Basile et al. (2019) dataset that consists of 10K English tweets labelled as either hateful (against immigrants or women) or not hateful. We use the unseen dataset to simulate how the models would perform in a realistic use case when a model is trained on one dataset and applied to a different dataset.

We want to study the effect of bias and how this affects the results when using our value-sensitive metric for evaluating the models with a reject option. We expect that the accuracy of the predictions on the unseen dataset is significantly lower than the accuracy of the predictions on the seen dataset, in line with the findings of related studies by Arango et al. (2019) and Gröndahl et al. (2018). Therefore, we also expect that the output value of our value-sensitive metric for the unseen dataset will be

[6] https://pypi.org/project/wordsegment/

lower and that the optimal rejection threshold will be higher (meaning that we need to reject more predictions).

3.3.5 Probability density functions

Since our value-sensitive rejector depends on the PDFs of the confidence values of the TP, TN, FP, and FN predictions, we need to empirically estimate these PDFs as we do not know the actual underlying distributions. We use the kernel density estimation (KDE) method provided by Statsmodels[7] for estimating these PDFs. With KDE, we estimate the PDF by weighing the confidence values from a set of predictions using a kernel function, a gaussian density function since it is the most commonly used, for each possible confidence value in the range $[0, 1]$. If there are many predictions with a confidence value around 0.8, then the KDE estimate will be higher around that point. The kernel function used in the KDE method also depends on a bandwidth (smoothness) value. A small bandwidth value results in an estimated PDF with much variance, while a high bandwidth value results in an estimated PDF with much bias. We use maximum likelihood cross-validation to find the optimal bandwidth value.

3.3.6 Application of the value-sensitive rejector

We apply the training phase of the value-sensitive rejector (refer to figure 3.1) to all three models for both the seen and the unseen datasets. Therefore, we use our metric from section 3.1 to calculate the total value $V(\tau)$ (formula 3.5) at all possible rejection thresholds (τ) for all different setups. We determine the optimal rejection threshold τ_O using the formulation from 3.6. Since we have a binary classification setting (hate or not hate), all confidence values will always be greater than or equal to 0.5. So if $\tau \in [0.0, 0.5]$, we accept all predictions and if $\tau = 1.0$, we reject all predictions. Therefore, we only calculate the total value of all predictions for the range $\tau \in [0.5, 1.0]$.

The first goal is to check the rejector's behaviour on different models and datasets. We can analyze this by plotting $V(\tau)$ for the range $\tau \in [0.5, 1.0]$, measuring the rejection rate (RR, percentage of rejected predictions), and measuring the accuracy of the accepted predictions. The second goal is determining whether the rejector can enhance hate speech detection. If the total value of a model for some optimal rejection threshold ($0.5 < \tau_O < 1.0$) is positive, then we know that the reject option can be beneficial for that specific model. The final goal is to compare the value-sensitive metric to machine metrics such as accuracy. We accomplish this by comparing the $V(\tau_O)$ values and the accuracies of all models.

[7] https://www.statsmodels.org/

Chapter 4

Survey study

The second part of this research is to find out how we can determine the value ratios between TP, TN, FP, FN, and rejected predictions. We conducted a literature study in section 2.5 and concluded that we want to empirically estimate the social values from the perspective of the social media user. In section 2.5.2, we found that ME is a promising technique for estimating subjective value ratios. Therefore, this chapter discusses how we apply the ME technique in a crowdsourced survey study.

We design a survey study to ask participants the degree to which they agree or disagree with the decisions of a fictional social media platform called SocialNet. We show the participants different scenarios representing TP, TN, FP, FN, and rejected predictions in the context of hate speech detection. The TP and TN scenarios mean that SocialNet successfully detects whether a post is hateful or not, respectively. The FP scenario means that SocialNet incorrectly predicts a non-hateful post as hateful, while the FN scenario implies that SocialNet incorrectly predicts a hateful post as non-hateful. For example, in the FN scenario, the survey shows a hateful post to the participant and explains that SocialNet did not identify the post as hate speech. Then, participants indicate their degree of agreement/disagreement using a scale, and we aggregate the answers per scenario type to obtain the value ratios.

The structure and preparation of our crowdsourced survey study follow the pre-registration plan for social psychology suggested by Van't Veer and Giner-Sorolla (2016). In a pre-registration plan, we describe the hypothesis, procedure, and analysis before conducting the crowdsourced survey study to increase scientific credibility, increase reproducibility and reduce bias (Van't Veer & Giner-Sorolla, 2016). It is essential to select the statistical methods for the analysis part beforehand to prevent ourselves from selecting the statistic that best fits the collected data. The content of this chapter reflects the final version of the pre-registration plan created after conducting the pilot survey.

In section 4.1, we make a hypothesis about the ME method and the value ratios. Section 4.2 contains all details about the survey setup. Finally, in section 4.3, we elaborate on the analysis of the survey results. Refer to Appendix C for the source code of the experimental setup and analysis of our survey study.

4.1 Hypotheses

We listed several hypotheses about the value ratios and the ME method before we conducted the survey experiment. The goal is to reflect on the hypotheses in the discussion to explain why specific results were expected or unexpected.

- **We hypothesize that the values of FP and FN are negative and that the value of an FN is lower than an FP**. We believe that both FP and FN predictions harm social media users; therefore, we think both values should be negative.

We believe that allowing hateful content to be publicly visible does more harm to social media users than filtering out neutral content. Therefore, we think an FN's value is lower than an FP's.

- **We hypothesize that the values of TP and TN are both positive and that the value of a TP is greater than a TN.** We believe that both TP and TN predictions positively impact social media users; therefore, we think both values should be positive. We believe predicting hateful content correctly is more valuable to social media users than correctly predicting non-hateful content. Therefore, we think a TP's value is greater than a TN's.
- **We hypothesize that the rejection value is negative and greater than the average value of an FP and an FN.** The critical assumption of using ML models with a reject option is that the negative value of rejection should always be greater than the negative value of an incorrect decision. Otherwise, rejecting predictions serves no purpose.
- **We hypothesize that FP and FN's absolute magnitudes are greater than TP and TN's.** We believe that social media users find the harm of incorrect predictions more critical than the benefits of correct predictions.
- **We hypothesize that ME is a suitable technique for retrieving the value ratios.** ME seems like a promising technique for retrieving ratio data from judgements about hate speech detection scenarios. We use a 100-level numerical scale for validation. We expect that both scales are correlated and will give similar judgements. Although we also expect the 100-level scale to be suitable for retrieving opinions about the different hate speech detection scenarios, it does not provide the ratio data we need. We also expect that the inter-rater reliability for the 100-level scale will be higher than for the ME scale since the ME scale provides more response freedom. We also expect this since the authors of Roitero et al. (2018) concluded that the inter-rater reliability of the 100-level scale is higher than the ME scale when rating the relevance of documents.

4.2 Method

This section discusses the complete setup of the survey experiment and how we use both scales.

4.2.1 Scales

We use ME as the primary scale of our survey experiment. As we concluded in section 2.5.2, we must also validate the ME scale. We validate the ME scale through cross-modality validation by comparing the results of the ME scale with another scale, as explained in section 2.5.2. The secondary scale is a bounded scale of 100 levels, called the 100-level scale, and we use this scale for four reasons. First, given the limited budget, it is impractical in this project to use other ME scales, such as measuring the intensity of the participants' handgrips to express their judgements. Second, there is no suitable dataset we can use for validation that contains human ratings of different scenarios in hate speech detection. Third, we concluded in 2.5.2 that Likert scales have limited response freedom. Finally, in Roitero et al. (2018), the authors concluded that the 100-level scale has several advantages over ME in terms of usability and reliability. The 100-level scale is easier to understand than ME, does

not require normalization, and provides more flexibility than a Likert scale (Roitero et al., 2018). Therefore, we create two separate surveys with the same scenarios where half of all participants use the 100-level scale and the other half use the ME scale. Both scales are bipolar scales since the participants should be able to either disagree or agree with the scenarios.

4.2.2 Normalization

The ME scale is unbounded and, therefore, provides a lot of response freedom. For example, suppose we first show a scenario, and the participant provides a value (e.g., 100) to indicate the degree of agreement. Suppose we next present a scenario that the participant agrees with more. The participant can always provide a higher value (e.g., 125). However, the results need to be normalized as different participants rate the agreement/disagreement degree differently. As explained in section 2.5.2, we cannot use standard normalization methods such as geometric averaging as we use bipolar scales with negative values. Therefore, we normalize the results by dividing the magnitude estimates of each participant by their maximum estimate. We multiply the normalized magnitude estimates by 100 for the sake of clarity. This way, all magnitude estimates are in the range $[-100, 100]$ while maintaining the ratio properties.

4.2.3 Design

This section lists all independent, dependent, confounding, and control variables analyzed in our experiment.

Independent variables

Independent variables are the different hate speech detection scenarios we show to the participants (TP, TN, FP, FN, and rejection). We inform the participants in the case of TP and FP scenarios that SocialNet ranks the hateful post lower on their feed. The users then need to spend more effort finding the post since they need to scroll longer before it becomes visible.

Initially, in the pilot survey, we explained that detected hateful posts are removed, which could be controversial. Also, we found that participants agreed more with the TP and TN scenarios compared to the degree to which they disagreed with the FP and FN scenarios. Therefore, we decided to explain that hateful posts are ranked lower, that it is expected from detection systems to produce correct predictions, and that incorrect predictions might cause harm to social media users. We did this to prepare the participants to focus on evaluating harm (instead of giving rewards).

We inform participants in the rejection scenarios that a human moderator needs to check the post (that can be either hateful or not hateful) within 24 hours. Meanwhile, the post remains visible with its original rank on the user's feed. We use 24 hours based on the German NetzDG law, which allows the government to fine social media platforms if they do not remove illegal hate speech within 24 hours (Tworek & Leerssen, 2019).

- **True Positive** Show a hateful post to the user and explain that SocialNet detected hate and ranked the post lower on people's feeds.
- **True Negative** Show a non-hateful post to the user and explain that SocialNet did not detect hate and allowed the post.

- **False Positive** Show a non-hateful post to the user and explain that SocialNet detected hate and ranked the post lower on people's feeds.
- **False Negative** Show a hateful post to the users and explain that SocialNet did not detect hate and allowed the post.
- **Rejection**
  - Show a hateful post to the user and explain that SocialNet was uncertain whether the post was hateful or not. An internal moderator will need to check the post within 24 hours. Meanwhile, the post remains visible.
  - Or show a non-hateful post to the user and explain that SocialNet was uncertain whether the post was hateful or not. An internal moderator will need to check the post within 24 hours. Meanwhile, the post remains visible.

Confounding variables

Confounding variables are the different demographic characteristics:

- **Nationality** People from different nationalities might have different perceptions and definitions of hate speech and opinions about how we should deal with it.
- **Ethnicity** People from different ethnicities might have different perceptions and definitions of hate speech and opinions about how we should deal with it.
- **Age** People of different ages might have different perceptions and definitions of hate speech and opinions about how we should deal with it.
- **Education** People with different educational statuses might have different perceptions and definitions of hate speech and opinions about how we should deal with it.
- **Sex** According to Gold and Zesch (2018), there is no significant difference in how men and women perceive hate. However, we still report sex as a confounding variable since we want to analyze if there are genuinely not any differences.

Control variables

We define two control variables: the measurement scales and the content of the social media posts we show to the participants. We control the measurement scale variable by randomly assigning a participant to use either the 100-level or the ME scale to rate the scenarios. Regarding the scales, as described before, we choose ME as our primary scale and use the 100-level scale for validation. We leave the study of other scales to future work. We control the content of the social media posts in two manners. First, we present all scenarios for all participants randomly to reduce bias. Second, we sample the social media posts for the survey from existing datasets. We explain the selection procedure in section 4.2.5.

- **Scales** The first group of participants must answer the questions using the ME scale. The second group needs to answer the questions using the 100-level scale.

- **Content of the posts** We sample all social media posts from existing datasets and present them to the participants in random order.

Dependent variables

Our dependent variables are the response values, reliability, validity and the value ratio of TP, TN, FP, FN, and rejection scenarios. Refer to section 4.3 for the reasoning and calculation of the following dependent variables.

- **Response values** All response values the participants give to the different scenarios with either the ME or the 100-level scale.
- **Reliability** The inter-rater reliability measured using Krippendorff's alpha, where values larger than 0.8 indicate reliable conclusions, and values larger than 0.6 indicate tentative conclusions (Krippendorff, 2004).
- **Validity** Convergent validity, if two different measures measure the same thing (Fitzner, 2007). Measured by calculating the correlation between the magnitude estimates and the response values from the 100-level scale.
- **Value ratios of TP, TN, FP, FN, and rejection scenarios** Measured by calculating the median of the normalized magnitude estimate response values of each scenario question and then calculating the mean over the resulting values to come up with the final value for that scenario type.

4.2.4 Planned sample

This section discusses how we pick the sample size and recruit the participants, and this section explains which stopping and exclusion rules we apply.

Sample size

There are 4.55 billion active social media users[1]. We choose a 90% confidence interval and 10% margin of error (MOE) for this study. So 90% of the time, our observations will fall within a 10% interval (Olson & Kellogg, 2014). According to Olson and Kellogg (2014), we need a sample size of 68 participants per survey type to reach the desired confidence interval and MOE values. We choose 10% MOE since we have a limited budget. We first conduct a pilot survey for 12 participants per scale to gather feedback and check if we need to improve things before the actual experiment. We want to determine the average workload using the pilot survey and whether reducing the MOE by increasing the number of participants is possible. For the pilot survey, we use 24 participants. Therefore, in total we need $2 * 12 + 2 * 68 = 160$ participants. Of the recruited participants, 50% identified as female. Half of the participants are assigned the ME scale, and the other half the 100-level scale.

Participants

We use the Prolific platform for recruiting online participants for the survey study. We use the following inclusion criteria for our participants:

- 18 years of age and older since we show offensive language in the experiment.

[1] https://datareportal.com/reports/digital-2021-october-global-statshot

- Fluent in English.
- Approval rating over 90% on the Prolific platform.
- Use one of the following social media platforms regularly (at least once a month): Facebook, Twitter, YouTube, LinkedIn, Pinterest, Google Plus, Tumblr, Instagram, Reddit, VK, Flickr, Vine.co, Meetup, ask.fm, Snapchat, TikTok, Medium.

Every participant is paid based on the hourly wage of 9.0 GBP (about 10,67 Euro), indicated as good pay by the platform[2]. We use the following exclusion/rejection criteria:

- Participants who fail the two attention checks. We include two instructional manipulation checks to check if the user pays attention to the survey[3].
- Participants who do not complete all questions.
- Participants who disagree with the informed consent before the start of the survey. We are not allowed to collect and process their data if they do not consent.

We select a balanced set of participants in Prolific, among which 50% are men and 50% are women.

4.2.5 Data

Depending on the assigned survey group, all subjects must judge several TP, TN, FP, FN, and rejection scenarios using either the ME or the 100-level scale. We select the posts used in the scenarios from a public dataset (Basile et al., 2019) that contains 13,000 English tweets. Each tweet is annotated with three categories: hate speech (yes/no), target (generic group or an individual), and aggressiveness (yes/no). Therefore, we have one neutral and four groups of hateful tweets: generic target + aggressive, individual target + aggressive, generic target + non-aggressive, and individual target + aggressive. For the rejection scenarios, we need both neutral and hateful tweets. Therefore, we need at least eight tweets per scenario type (TP, TN, FP, FN, and rejection). We need 40 tweets, where 20 are hateful, and 20 are not hateful, to create 40 different scenarios.

We want to select the most representative tweets from the dataset. Randomly selecting the tweets from the dataset is insufficient as the dataset might contain sample retrieval bias, as explained in section 2.1. We might retrieve too many similar tweets about the same topic when randomly selecting the tweets. Therefore, we perform content analysis to create a selection of tweets that is as representative and diverse as possible. We provide an overview of our selection process in figure 4.1.

We exclude all tweets that contain Twitter replies and mentions since they have unclear contexts. Then we preprocess all tweets by removing the URLs and hashtags. Finally, we use clustering analysis to select 40 tweets for our study. We perform latent semantic analysis (LSA) and k-means clustering on each group of tweets.

We use the term frequency-inverse document frequency (TF-IDF) to represent all documents and their words, also known as terms, in a matrix where the term frequencies indicate how important that term is to the document (Aggarwal & Zhai,

[2] https://prolific.co/pricing
[3] https://researcher-help.prolific.co/hc/en-gb/articles/360009223553

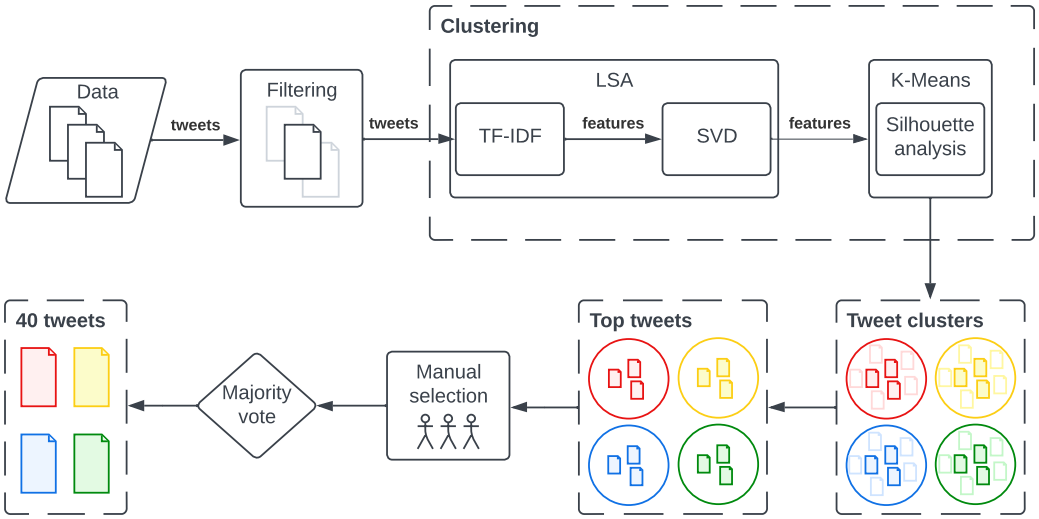


FIGURE 4.1: Flow diagram that visualizes how we perform content analysis to cluster and select the tweets for our survey study.

2012). The term frequencies are multiplied by the inverse document frequency so that terms that often occur in all documents, such as stop words, will end up with a lower value in the matrix (Aggarwal & Zhai, 2012).

Then, we use singular value decomposition (SVD) for dimensionality reduction to transform the output matrix of the TF-IDF step. The transformed matrix is more suitable for text clustering techniques since documents with similar terms are now grouped (Aggarwal & Zhai, 2012). The combination of TF-IDF and SVD is also known as LSA and is suitable for clustering purposes (Aggarwal & Zhai, 2012).

Finally, we apply the unsupervised learning technique k-means to the output of the LSA method to cluster all tweets into k clusters. We calculate the silhouette coefficient to determine the optimal cluster size (k value) for the neutral tweets and the four groups of hateful tweets. The silhouette analysis indicates setting k as large as possible.

We select the five nearest data samples to each cluster centroid. From this selection, we manually choose one tweet per cluster using a majority vote from three group members to create the final set of 40 tweets. Based on the silhouette coefficient, we use a cluster size of 20 for the neutral tweets and select one tweet per cluster to collect 20 neutral tweets. Furthermore, we use a cluster size of 5 for each group of hateful tweets to collect 20 hateful tweets.

Refer to appendix A.1 for the resulting list of all scenarios.

4.2.6 Procedure

In figure 4.2, we present the procedure of the two surveys, one where participants use the ME scale and another where participants use the 100-level scale. We use LimeSurvey[4] as our survey tool. The survey first presents the informed consent policy and excludes participants that do not agree with it. Next, we show introductory texts to the participants to explain what we expect from them and to explain the structure of the survey. Using the ME scale, we first present a training phase where the participants need to estimate five different line lengths using any positive

[4] https://www.limesurvey.org/

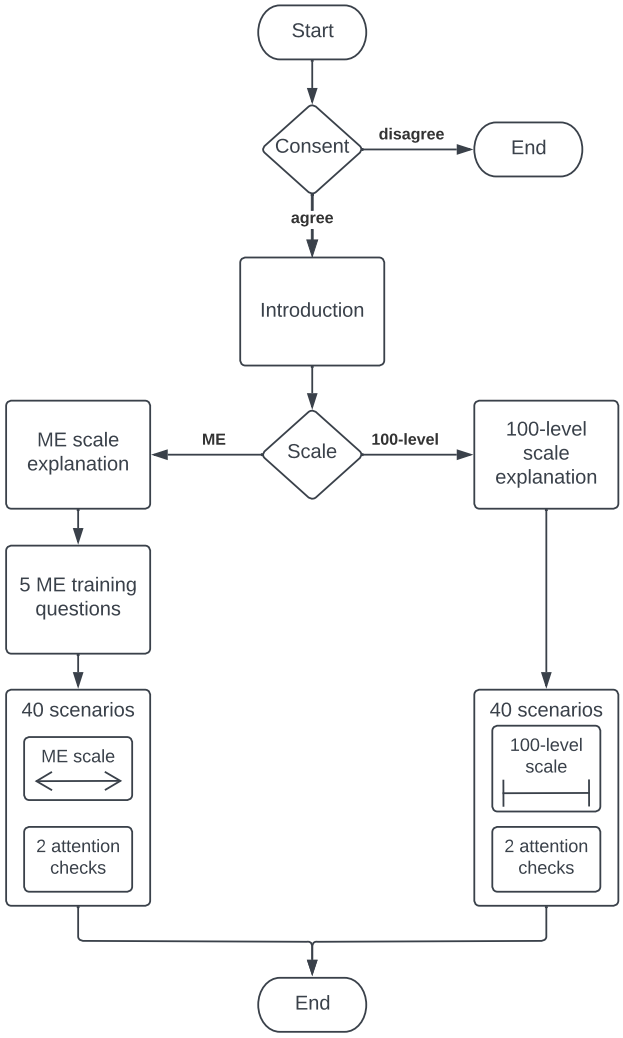


FIGURE 4.2: Flow diagram that visualizes the procedure of the survey study. We assign half of the participants to the ME survey and the other half to the 100-level survey.

value to get familiar with using the ME scale. Then, we randomly present two attention checks and 40 scenarios representing the TP, TN, FP, FN, and rejection scenarios (with eight scenarios per type). Each scenario contains several questions with the same structure. The first question is whether participants think the post is hateful (yes/no). The second question is whether participants agree, disagree, or are neutral with SocialNet's decision. In the case of nonneutral, we ask a third question about the degree to which participants agree or disagree with the machine's decisions, using either the ME or 100-level scale, depending on their group. There is no time limit for answering the questions, and all data is anonymous. Finally, we inform the participants not to put personal identifiers in their answers. Refer to Appendix A for all presentation texts, the informed consent, and some scenario examples.

4.3 Analysis

First, we calculate the value ratios between the TP, TN, FP, FN, and rejection scenarios in hate speech detection using the survey's results. Second, we analyze the quality of our survey method by looking at two aspects: reliability and validity.

4.3.1 Value ratios

The survey study aims to determine the value ratios between the TP, TN, FP, FN, and rejection scenarios in the context of hate speech detection. The metric from section 3.1 takes these numerical values as input to calculate the optimal rejection threshold. We do not need to know the absolute values but only the relative values. For example, if we set all values to 1, we retrieve the same optimal rejection threshold as setting all values to 1000. We use a bipolar scale for question 3 in the survey since we ask the participants the degree to which they agree, disagree, or are neutral with the decision of SocialNet. For both scales, we convert disagreement values to negative values, neutral values to 0, and agreement values to positive values. Since we found that the data of both scales is skewed after conducting the pilot survey, we first apply the median to the individual questions' results. Then we calculate the mean value over the resulting values to retrieve the final aggregated value ratios. For example, to calculate the aggregated V_{tp} values for both scales, we use:

$$V_{tp}^{ME} = \frac{1}{n}\sum_{i=1}^{n} \tilde{r}_{i,tp}^{ME}$$

where n is the total number of TP scenarios, and $\tilde{r}_{i,tp}^{ME}$ is the median response value of TP question number i rated by all participants with the ME scale.

$$V_{tp}^{100L} = \frac{1}{n}\sum_{i=1}^{n} \tilde{r}_{i,tp}^{100L}$$

where n is the total number of TP scenarios, and $\tilde{r}_{i,tp}^{100L}$ is the median response value of TP question number i rated by all participants with the 100-level scale.

We apply the same calculations for the remaining scenario types. The results should give us an understanding of how the participants feel towards the different scenarios: TP, TN, FP, FP, and rejection. We define the value ratios we need for the metric using the aggregated values of the TP, TN, FP, FN, and rejection scenarios rated with the ME scale since the ME scale provides us with ratio data. We do not use the aggregated values of the 100-level scale for our metric since the 100-level scale does not provide ratio data, but we still present them.

4.3.2 Reliability

Reliability is about whether we can trust our results and if we get consistent results (Fitzner, 2007). We do this by mainly looking at inter-rater reliability. Different participants should give approximately the same judgements to the same scenarios. We measure the inter-rater reliability using Krippendorff's alpha (Krippendorff, 2004; Maddalena et al., 2017). We calculate the inter-rater reliability value for the complete survey's data for the normalized ME and 100-level values. We use the inter-rater reliability scores to compare the ME scale with the 100-level scale. We also separately study the inter-rater reliability values for the different types of scenarios (TP, TN, FN, FP, and rejection). This experiment does not consider other types of reliability, such as test-retest reliability. Guaranteeing test-retest reliability would require us to redo the complete experiment at a different time for the same participants, which is infeasible for this project, given the limited time and budget.

4.3.3 Validity

Validity is about whether we are measuring the things we want to measure (Fitzner, 2007). The main goal of this aspect is to validate if we can use the ME technique to measure participants' opinions about hate speech detection scenarios. There are multiple types of validity, but we focus mainly on convergent validity (part of construct validity), content validity, and face validity (Fitzner, 2007). Construct validity checks whether there is an agreement between a theory and a measurement device or procedure (Fitzner, 2007). Convergent validity is about the correlation between different measures to see if they measure the same phenomenon (Fitzner, 2007). Content validity is about letting experts review the proposed research questions and procedure (Fitzner, 2007). Face validity is a subjective type of validity, and it is about why we think the questions and proposed procedures are valid (Fitzner, 2007).

We analyze convergent validity by performing cross-modality validation. Following the approach from Roitero et al. (2018), we analyze the correlation between the ME scale and the 100-level scale. We can verify that they measure the same phenomenon if we find that both scales are positively correlated. However, we can also expect a low correlation since the ME scale is a (normalized) unbounded scale, and the 100-level scale is bounded. Nevertheless, we think both scales give similar results, meaning that high ME responses should correspond to high 100-level scale responses and low ME responses to low 100-level scale responses. To guarantee content validity, we let experts (the supervisors of this thesis project) check the pre-registration report before conducting the experiments. We tackled face validity in section 2.5 by arguing why we think the ME technique is suitable for measuring people's opinions about hate speech detection scenarios. We exclude other forms of validity from this experiment because they are irrelevant or infeasible. For example, external validity is about the degree to which the findings can be generalized to other settings or groups (Fitzner, 2007). We think people with different demographic characteristics perceive hate speech differently since people have other norms and values. We believe that if we conduct this experiment using different groups of participants, we might retrieve different value ratios. Therefore, we decided not to create too many participant inclusion criteria but take a random sample of global social media users. We would have to experiment with multiple groups with different demographic characteristics to analyze external validity. We left this for future work to investigate in full detail. However, we still try to analyze if we can find any differences between participants with different demographic characteristics in the dataset we retrieve (refer to section 4.3.4).

4.3.4 Demographics

As we conduct the survey study only once for a group of participants, among which 50% are men and 50% are women, the remaining demographic characteristics can be quite diverse. Nevertheless, we verify whether there are any significant statistical differences between groups of participants with different demographic characteristics. We expect that demographic characteristics influence people's perception of hate speech and how we should deal with it. Therefore, we apply several statistics to the results of each scenario to analyze if we can find differences between different demographic groups.

Prolific provides information about the demographic characteristics of the participants, out of which we analyze six features: sex, student (whether they are still a student or not), continent, nationality, language, and ethnicity. We manually add

the continent feature based on the values of the nationality feature. Most features overlap with our pre-defined confounding variables from section 4.2.3, where features such as nationality, continent, and language are highly correlated. We exclude age as almost all participants fall between 20 and 30 years old.

We have multiple groups (more than two) for nationality, ethnicity, and language and two groups for the features student, sex, and continent (since we found only two continents in the demographic data of all participants). We apply either analysis of variance (ANOVA) (parametric) or Kruskal-Wallis (non-parametric) when we have more than two groups. Furthermore, we apply an unpaired two-sample t-test (parametric) or the Mann-Whitney U Test (non-parametric) when we have exactly two groups.

First, we check if we can apply the parametric statistics by checking if their assumptions hold in our dataset. If not, then we use the non-parametric tests. We apply ANOVA and the t-test when the data meets the following three conditions: homogeneity of variance (each population has the same variance), normality (the data of each population is normally distributed), and independence (the observations are independent of each other) (Howell, 2012). We use Bartlett's test of homogeneity of variances and the Shapiro-Wilk test of normality to check if we can apply ANOVA and the t-test. We obey the independence condition since we collect the data of all participants independently.

ANOVA and the t-test can be robust to violations of the homogeneity of variances and the normality assumptions (Howell, 2012). However, if one of the assumptions is violated, then it is essential to keep the sample sizes as equal as possible (Howell, 2012).

Finally, for the multi-group features (nationality, language, and ethnicity), we apply pairwise statistical tests (Mann-Whitney U or t-test) between all groups. We only do this for the scenarios where we find significant differences between the groups through ANOVA/Kruskal-Wallis.

However, we now may introduce Type I errors as we perform many pairwise statistical tests between all groups. As a result, we might incorrectly reject the null hypothesis for some pairwise tests, meaning that we find significant differences between some groups while there are none. Therefore, we perform the post hoc Benjamini-Hochberg procedure to correct the p values of the pairwise test results to control the Type I errors.

Chapter 5

Results

This chapter presents the results of the survey study (chapter 4) and the experiments with the value-sensitive rejector (chapter 3). We first present the results of the survey study as the experiments with the value-sensitive rejector depend on the outcomes of the survey study.

The goal of the survey study was to retrieve the value ratios of TP, TN, FP, FN, and rejected predictions in hate speech detection from the perspective of the social media user. We retrieved the value ratios using the ME scale. We validated the ME scale by conducting a separate survey using a bounded scale of 100 levels, called the 100-level scale.

We defined three goals of the experiments with the value-sensitive rejector. First, we want to analyze how the rejector behaves on different models and datasets. Second, we want to find out if rejecting predictions increases the utilities of the ML models in terms of the value of our value-sensitive metric. Finally, we want to compare the value-sensitive metric against machine metrics such as accuracy.

Section 5.1 covers the results of the complete survey study that we collected after conducting the pilot survey, and section 5.2 covers the results of the experiments with the value-sensitive rejector.

5.1 Survey study

We collected the responses of all participants to all scenarios for both surveys: one group that uses the ME scale and another that uses the 100-level scale. All participants had to answer two/three questions per scenario, dependent on the choice of the second question.

The first question asked whether the participant found the content of the social media post hateful or not. Figure 5.1 presents the results of the first question by showing the percentages of participants who find the content hateful or not hateful for each scenario. We summed the ME and the 100-level survey responses since this question was the same for both surveys. Most participants agreed with the ground truth label of the social media posts. Please note that according to the ground truth label, REJ1, REJ2, REJ5, and REJ6 are hateful, and REJ3, REJ4, REJ7, and REJ8 are not hateful. So most participants found the posts used in the TP and FN scenarios hateful, and those used in TN and FP scenarios not hateful. For the rejection scenarios, most found the posts of REJ1, REJ2, and REJ6 hateful and REJ3, REJ4, and REJ8 not hateful. However, we found three posts where a significant number of participants tended to disagree with the ground truth label (more than or equal to 40%): FN5, REJ5, and REJ7.

The second and third questions asked whether the participant agreed/disagreed or was neutral about SocialNet's decision and to what degree. Figure 5.2 shows the response values to the second and third questions of all scenarios for both scales.

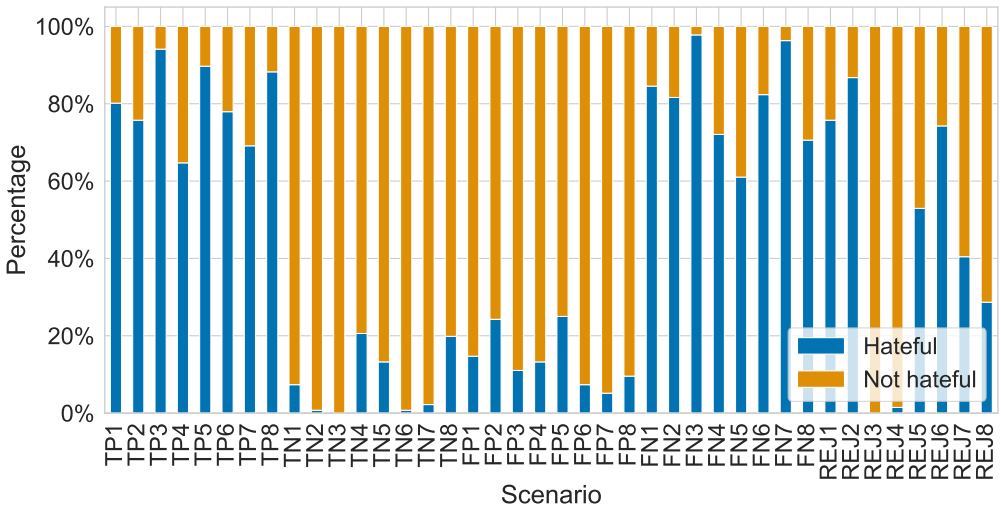


FIGURE 5.1: Stacked bar charts that show the percentages of participants who find the content of the social media post used in the scenarios hateful or not hateful. Each bar is a summation of the responses to both surveys, as this question was the same for both.

Participants generally agreed with the TP and TN scenarios and disagreed with the FP, FN, and rejection scenarios. For both scales, participants disagreed the most with scenarios FN3 and FN7 and agreed the most with scenarios TN3 and TN6.

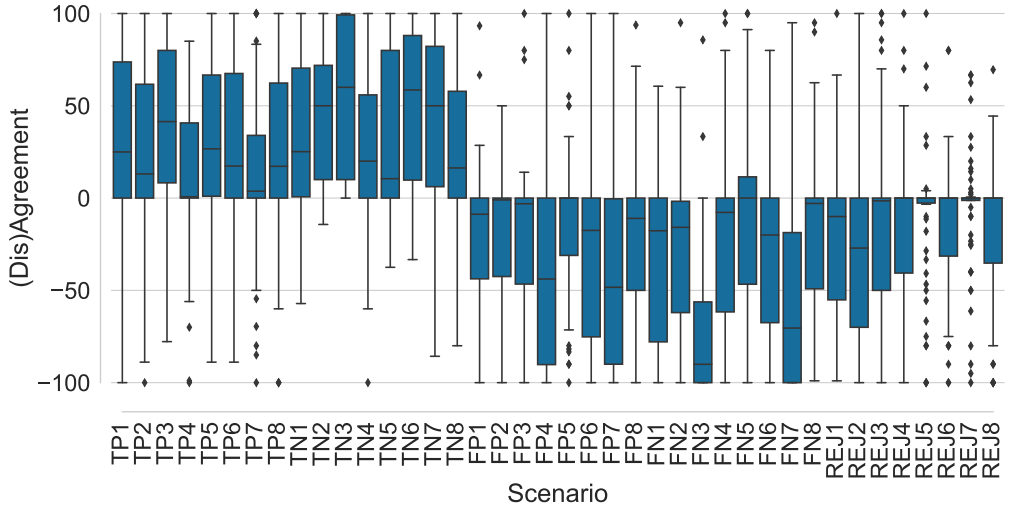
The following sections tackle the different parts of the survey analysis: section 5.1.1 presents the value ratios required for our value-sensitive rejector, section 5.1.2 presents the reliability analysis, and section 5.1.3 the validity analysis. Finally, section 5.1.4 shows the results of the demographic analysis.

5.1.1 Value ratios

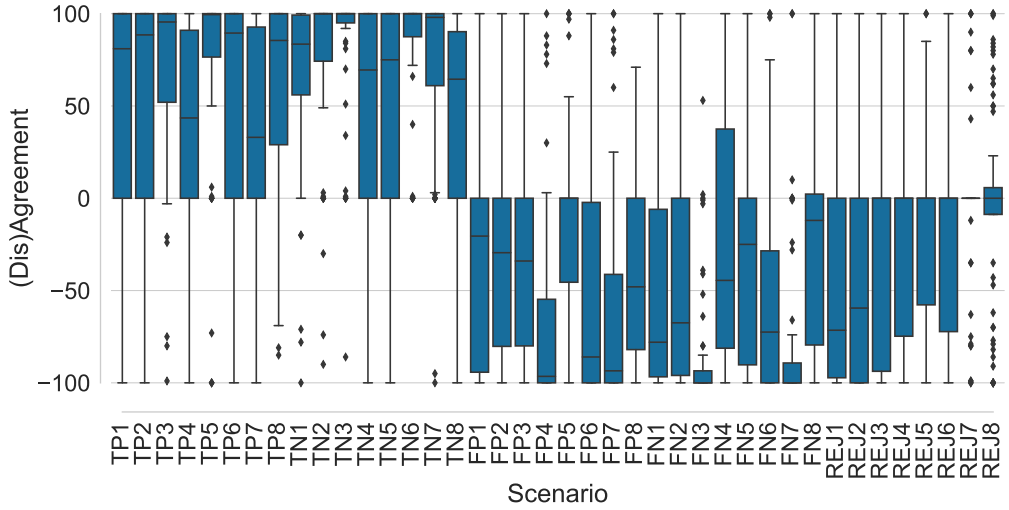
We need the value ratios between TP, TN, FP, FN, and rejected predictions in the context of hate speech detection to use our value-sensitive rejector from chapter 3 for calculating the optimal rejection threshold. We calculated the value ratios following the approach from section 4.3.1. Table 5.1 shows the resulting values (v) from the ME and the 100-level surveys. Positive and negative values indicate agreement and disagreement, respectively. For both scales, participants disagreed the most with the FN scenarios and agreed the most with the TN scenarios. The final values of both scales follow the same order: $V_{fn} < V_{fp} < V_r < V_{tp} < V_{tn}$. Participants gave the highest absolute response values to the TN scenarios. We also observed that participants provided greater absolute response values to the TP and TN scenarios than to the FP and FN scenarios.

5.1.2 Reliability

As explained in section 4.3.2, we measured the interrater reliability between the participants using Krippendorff's alpha. Table 5.1 shows Krippendorff's alpha (α) values for both scales. In the last row of the table, we computed the α values over the responses to all scenarios. The ME scale seemed more reliable than the 100-level scale. According to Krippendorff (2004), the results of the ME scale are reliable, while the results of the 100-level scale are likely to be unreliable.



(A) ME scale



(B) 100-level scale

FIGURE 5.2: Boxplots of the responses of all participants to all scenarios for both scales.

We also computed the α values for each group of scenarios of the same type (TP, TN, FP, FN, or rejection). Participants using the ME scale tended to agree with each other on the FP and FN scenarios, while they tended to disagree on the rejection scenarios. For the 100-level scale, we see that participants have low agreement on all scenario types.

5.1.3 Validity

We analyzed the validity of the ME method by performing cross-modality validation between the ME and the 100-level scale (refer to section 4.3.3). Figure 5.3 shows the correlation between the ME scale and the 100-level scale. The Shapiro-Wilk test of normality showed that both the median (normalized) ME scores and the median 100-level scores do not follow a normal distribution ($p < 0.05$). We calculated the median because when we look at figure 5.2, we can see that the data of both scales are skewed and contain many extreme outliers. We calculated the Spearman and

	ME		100-level	
	α	v	α	v
TP	0.07	18.15	0.04	77.00
TN	0.10	36.32	0.11	86.31
FP	0.39	-16.69	0.07	-51.00
FN	0.92	-28.08	0.14	-62.43
Rejection	-0.31	-4.82	0.07	-16.37
All	0.78	—	0.44	—

TABLE 5.1: Krippendorff's alpha (α) and the scenario values (v) for TP, TN, FP, FN, and rejection scenarios for the ME and 100-level scales.

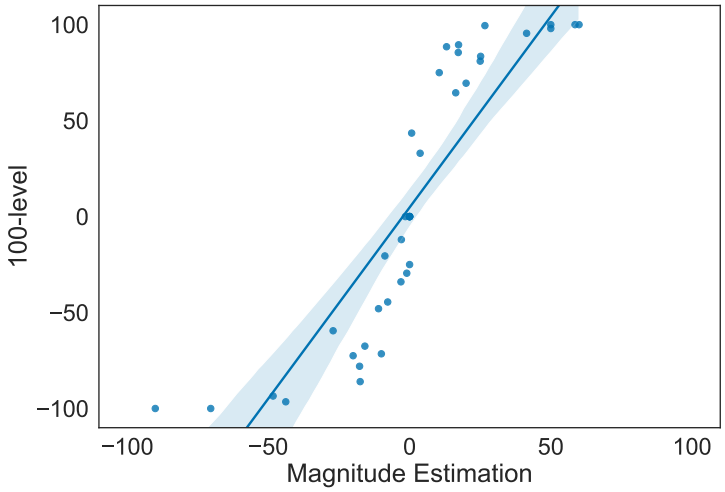


FIGURE 5.3: Correlation plot between the median normalized magnitude estimates and the median 100-level scores per question.

the Kendall correlation statistics as these are non-parametric and, therefore, do not require the normality assumption. Spearman returned a 0.98 and Kendall a 0.89 correlation between the ME and the 100-level scales ($p < 0.05$), indicating that both scales are highly correlated.

5.1.4 Demographics

We followed the approach from section 4.3.4 to analyze whether statistically significant differences exist between groups with different demographic characteristics. We focused on six features: sex, student, continent, nationality, language, and ethnicity.

We only used non-parametric statistical tests to analyze the demographic differences for two reasons. First, we found that the assumptions of normality and homogeneity of variances were violated in our dataset when looking at the different groups for all features. Second, we found that for most features, except sex, the sample sizes of the feature groups were not equal. We used Mann-Whitney U to verify a significant difference between the two groups, and we used Kruskal-Wallis for more than two groups.

Table 5.2 shows the resulting p values for all scenarios and all features. We found that there are no significant differences between men and women. We found only

three scenarios with significant differences for the student and continent features. We found the most significant differences when looking at nationality and language. We observed five scenarios with non-hateful posts and ten scenarios with hateful posts where there is at least one feature with significant differences between the groups. Scenarios FP7 and REJ4 have the most features (4) with significant differences. Scenarios TP6, FN5, and REJ1 have the second most features (2) with significant differences.

Table B.1 shows the p values of the statistical tests for the aggregated scenarios (TP, TN, FP, FN, and rejection) and all features. We aggregated the scores by calculating the mean response value to all scenarios of the same type, e.g. TP, for each participant. Then we applied the statistical tests to the aggregated scores. We found the most significant differences in the aggregated scores of the FP scenarios.

Finally, we conducted pairwise Mann-Whitney U tests to check if there were any significant differences between pairs of groups for the multi-group features: nationality, language, and ethnicity. Tables B.2, B.3, and B.4 present the resulting p values of the pairwise Mann-Whitney U tests for the features of nationality, language, and ethnicity, respectively. We did not find many pairwise significant differences for most of these scenarios and the three features. We found the most pairwise differences (four out of the six pairs) for scenario FN5 and the nationality feature.

5.2 Value-sensitive rejection

We experimented with our value-sensitive rejector following the approach from section 3.3.6. We produced a set of predictions for each experimental setup, applied our value-sensitive rejector to each setup, and collected the results for analysis.

We used all three models (LR, DistilBERT, and CNN) to produce predictions for both the *seen* and *unseen* datasets. Therefore, we ended up with six different sets of predictions. Then, for each set of predictions, we created the PDFs using KDE for all predictions of the same type (TP, TN, FP, and FN). The PDFs were necessary for calculating the total value of the models with the reject option. Figures B.1 and B.2 show all PDFs for the *seen* and *unseen* datasets, respectively. We observed that all three models were more confident in their correct predictions (TP and TN) than their incorrect predictions (FP and FN) for both the *seen* and *unseen* datasets. All three models were also more confident in their correct predictions for the *seen* dataset than the *unseen* dataset. The CNN and LR models have similar PDFs and seem more calibrated since the PDFs of the correct predictions are skewed towards 1.0. In contrast, the PDFs of the incorrect predictions follow a more uniform distribution. The DistilBERT model is less calibrated than the other two models. We recognize this in the PDFs of the incorrect predictions in figures B.1 and B.2 of the DistilBERT model by looking at the large density values around the high confidence values.

We applied the value-sensitive metric from section 3.1 to the three models and the two datasets using the PDFs and the ME values ($V_{tp}, V_{tn}, V_{fp}, V_{fn}$, and V_r) from the survey. Figure 5.4 presents the total value of all models with the reject option ($V(\tau)$) for all possible rejection thresholds ($\tau \in [0.5, 1.0]$) and the ME values from table 5.1. The diamond-shaped markers indicate the optimal rejection threshold (τ_O) for which the model achieves the highest total value ($V(\tau_O)$). Positive $V(\tau)$ values indicate that the model for rejection threshold τ is valuable, and negative values indicate that the costs of incorrect accepted/rejected predictions exceed the gains of correct accepted/rejected predictions. For all models, we got $\tau_O \approx 0.5$, meaning that all models achieve the highest total value when all predictions are accepted.

	Two groups			More than two groups		
	Sex	**Student**	**Continent**	**Nationality**	**Language**	**Ethnicity**
TP1	0.506	0.371	0.982	0.095	0.117	0.108
TP2	0.268	0.201	0.387	0.300	0.330	0.464
TP3	0.680	0.276	0.577	0.160	**0.046**	0.138
TP4	0.756	0.441	0.774	0.137	0.175	0.568
TP5	0.392	**0.011**	0.387	0.152	0.106	0.341
TP6	0.260	0.097	0.682	**0.002**	**0.006**	0.215
TP7	0.342	0.730	0.059	0.241	0.400	0.238
TP8	0.495	**0.015**	0.246	0.568	0.387	0.190
TN1	0.430	0.480	0.554	0.307	0.260	0.449
TN2	0.567	0.382	0.633	0.595	0.716	0.833
TN3	0.393	0.866	0.766	0.443	0.298	0.432
TN4	0.104	0.171	0.059	0.245	0.251	0.201
TN5	0.290	0.199	0.964	0.304	0.177	0.296
TN6	0.521	0.510	0.608	0.815	0.748	0.600
TN7	0.224	0.878	**0.050**	0.108	0.223	0.314
TN8	0.191	0.417	0.327	0.168	0.761	0.872
FP1	0.270	0.545	0.065	0.093	0.333	0.174
FP2	0.337	0.114	0.155	**0.008**	0.164	0.195
FP3	0.561	0.509	0.889	0.793	0.725	0.205
FP4	0.278	0.860	0.908	0.267	0.186	0.344
FP5	0.847	0.445	0.220	0.269	0.554	0.194
FP6	0.774	0.266	0.555	0.758	0.409	0.486
FP7	0.391	0.784	**0.015**	**0.026**	**0.020**	**0.010**
FP8	0.624	0.837	0.681	0.544	0.225	0.705
FN1	0.337	0.213	0.317	0.261	0.668	0.558
FN2	0.791	0.928	0.759	0.967	0.974	0.823
FN3	0.990	0.752	0.480	0.504	0.455	0.182
FN4	0.511	0.573	0.450	0.549	0.856	0.965
FN5	0.306	0.467	0.802	**0.001**	**0.009**	0.349
FN6	0.109	0.113	0.928	**0.012**	0.084	0.436
FN7	0.871	0.677	0.093	0.107	**0.046**	0.148
FN8	0.776	**0.009**	0.819	0.949	0.363	0.117
REJ1	0.799	0.734	0.544	**0.021**	**0.012**	0.168
REJ2	0.644	0.202	0.741	0.295	0.258	0.749
REJ3	0.803	0.815	0.108	0.425	0.482	0.133
REJ4	0.985	1.000	**0.002**	**0.014**	**0.036**	**0.002**
REJ5	0.133	0.994	0.570	0.111	**0.036**	0.090
REJ6	0.244	0.195	0.716	0.061	0.166	0.664
REJ7	0.911	0.853	0.942	0.997	0.996	**0.020**
REJ8	0.157	0.167	0.944	0.901	0.741	0.108

TABLE 5.2: **Individual**: an overview of the statistical differences between different groups of participants for various demographic characteristics for each scenario in the ME survey. Each cell contains the p value of either the Mann-Whitney U test for two groups or the Kruskal-Wallis test for more than two groups. The grey cells with bold text indicate significant statistical differences between the groups for that feature and scenario type.

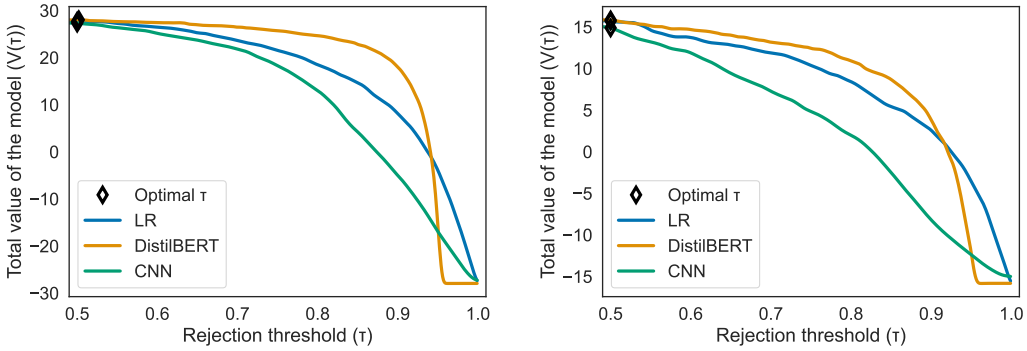
Figure 5.4 shows that all models' $V(\tau_O)$ values are greater for the *seen* data than for the *unseen* data. The total value of all models decreases for increasing values of the rejection threshold.

To further examine how $V(\tau)$ behaves when we only consider punishing incorrect predictions instead of rewarding correct predictions, we applied the metric again, setting V_{tp} and V_{tn} equal to zero. The metric's conditions 3.2b and 3.2a were still satisfied when we did this. Figure 5.5 presents the total value ($V(\tau)$) again for the updated values $V_{tp} = 0$ and $V_{tn} = 0$. We found that τ_O of all models moved towards 1.0, meaning that rejecting predictions is now more beneficial for the total value of the models. All models achieve the highest $V(\tau)$ value when $\tau \in [0.7, 0.9]$ for *seen* data and when $\tau \in [0.9, 1.0]$ for *unseen* data.

Table 5.3 shows the specific values of τ_O, the accuracies of the accepted predictions, and the rejection rates (fraction of rejected predictions). The first two rows show that the accuracies of all models dropped when we applied the models to *unseen* data. The last two rows (where $V_{tp} = 0$ and $V_{tn} = 0$) show that we achieved higher accuracies of accepted predictions for increasing optimal rejection thresholds. For all models, we rejected less than 30% of all predictions for the *seen* data and a large fraction for the *unseen* data. The DistilBERT model achieved the highest accuracies of accepted predictions for all configurations. For the *seen* data, it achieved an accuracy of accepted predictions of 92.6% while rejecting only 25.2% of all predictions. For the *unseen* data, it rejected the least amount of predictions (92.3%) and achieved the highest accuracy of accepted predictions (88.1%). The CNN model performs the worst for all configurations regarding the accuracy of accepted predictions. The CNN model achieved the highest value for the *unseen* data when all predictions were rejected, indicating that it is not valuable to use the CNN model.

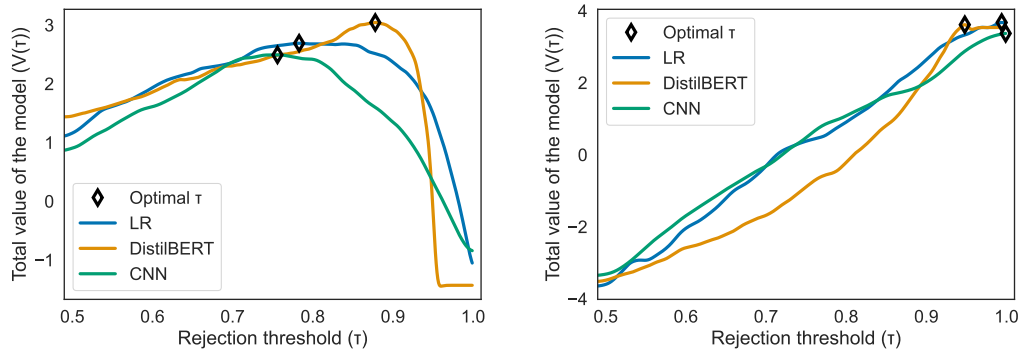
Table 5.4 compares the results of our value-sensitive metric with machine metrics like accuracy. For all models, it presents the total value for the optimal rejection thresholds ($V(\tau_O)$), the total value when all predictions are accepted ($V(0)$), and the accuracies when all predictions are accepted. First, we compared the accuracy of the original model with $V(0)$, as in both cases, all predictions were accepted. In the first two rows, both the accuracy and $V(0)$ indicate that the DistilBERT model performed the best for both the *seen* and *unseen* datasets. In the last two rows (where $V_{tp} = 0$ and $V_{tn} = 0$), both metrics indicate that the DistilBERT model performed the best for the *seen* dataset but got different results for the *unseen* dataset. For the *unseen* dataset, according to the accuracy, the DistilBERT model performed the best, while the CNN model performed the best according to $V(0)$. All $V(0)$ values in the last row show that none of the models is valuable for *unseen* data when we accept all predictions.

When we look at the $V(\tau_O)$ values in table 5.4, we see that all models are valuable for the optimal rejection threshold. The DistilBERT model achieved the highest $V(\tau_O)$ values in all configurations except for the *unseen* data with $V_{tp} = 0$ and $V_{tn} = 0$, as the LR model achieved a higher total value. This result is interesting as we can see from table 5.3 that for the DistilBERT model, the accuracy of the accepted predictions is higher, and the rejection rate is lower than for the LR model. By comparing $V(\tau_O)$ with $V(0)$ in the last row of the table, we can see that all models become valuable when we adopt the optimal rejection threshold.



(A) evaluated on *seen* data (B) evaluated on *unseen* data

FIGURE 5.4: $V(\tau)$ functions of all models with $V_{tp} = 18.15$, $V_{tn} = 36.32$, $V_{fp} = 16.69$, $V_{fn} = 28.08$, $V_r = 4.82$.



(A) evaluated on *seen* data (B) evaluated on *unseen* data

FIGURE 5.5: $V(\tau)$ functions of all models with $V_{tp} = 0.0$, $V_{tn} = 0.0$, $V_{fp} = 16.69$, $V_{fn} = 28.08$, $V_r = 4.82$.

	LR			DistilBERT			CNN		
	τ_O	Acc	RR	τ_O	Acc	RR	τ_O	Acc	RR
Seen data	0.500	0.847	0.000	0.502	0.850	0.000	0.500	0.835	0.000
Unseen data	0.500	0.640	0.000	0.500	0.640	0.000	0.500	0.629	0.000
Seen data ($V_{tp} = V_{tn} = 0$)	0.783	0.910	0.250	0.878	0.926	0.252	0.756	0.898	0.278
Unseen data ($V_{tp} = V_{tn} = 0$)	0.994	0.752	0.958	0.948	0.881	0.923	0.999	-	1.0

TABLE 5.3: The optimal rejection thresholds (τ_O), the accuracy of the accepted predictions (Acc), and the rejection rates (RR) of all models for both datasets.

	LR			DistilBERT			CNN		
	$V(\tau_O)$	$V(0)$	Acc	$V(\tau_O)$	$V(0)$	Acc	$V(\tau_O)$	$V(0)$	Acc
Seen data	27.707	27.707	0.847	28.001	27.996	0.850	27.291	27.291	0.835
Unseen data	15.689	15.689	0.640	15.823	15.823	0.640	14.868	14.868	0.629
Seen data ($V_{tp} = V_{tn} = 0$)	2.688	1.158	0.847	3.041	1.448	0.850	2.490	0.901	0.835
Unseen data ($V_{tp} = V_{tn} = 0$)	3.668	-3.605	0.640	3.606	-3.489	0.640	3.365	-3.322	0.629

TABLE 5.4: The maximum total values of the models for the optimal rejection threshold ($V(\tau_O)$), the total value of the models when all predictions are accepted ($V(0)$), and the accuracies (Acc) of all models.

Chapter 6

Discussion

The main goal of this project was to propose a way of rejecting ML model predictions in a value-sensitive manner for hate speech detection. We split this up into two parts. First, we wanted to determine how we could measure the total value of ML models with a reject option. We proposed a value-sensitive metric that measures the total value of an ML model for some rejection threshold, where we reject and accept all predictions with a confidence value below and above the threshold, respectively. This calculation is based on a set of predictions and the value ratios between TP, TN, FP, FN, and rejected predictions. By maximizing the total value, we can find the optimal rejection threshold. Second, we wanted to develop a method for determining these value ratios necessary for the metric's calculations. We proposed to estimate the value ratios in a large survey study using the ME scale. The results from chapter 5 showed several key findings:

- The survey's results indicated that the ME technique is suitable for retrieving the value ratios from human ratings to hate speech detection scenarios since the results passed both the reliability and validity analyses.
- Social media users appreciate correct predictions more than they detest incorrect predictions, especially content correctly identified as non-hateful and, therefore, not banned from the platform.
- Social media users agree more on rating the negative value of incorrect predictions than the positive value of correct predictions.
- We did not observe significant statistical differences for most scenarios between groups of participants with different demographical characteristics.
- The experiments with the *seen* data showed that our value-sensitive rejector maximizes the utility of hate speech detection models in terms of the value of our value-sensitive metric when we consider not rewarding correct predictions.
- The experiments with the *unseen* data demonstrated that hate speech detection models are susceptible to bias, affecting the results of our value-sensitive rejector since we had to reject more predictions when we considered not rewarding correct predictions. Also, the results showed that when using our value-sensitive metric, the best model selected can be different compared to using accuracy.

This chapter analyzes the results from chapter 5 in greater detail. First, we discuss the main findings of the survey study in section 6.1 and our value-sensitive rejector in section 6.2. Finally, we highlight some limitations of our approach in section 6.4 and give some recommendations in section 6.5.

6.1 Survey study

In each scenario, we first asked the participant to indicate if they thought the social media post was hateful or not. We found three scenarios for which more than 40% disagreed with the ground truth label (FN5, REJ5, and REJ7). We also recognized this in the ME response values since participants generally were neutral about these scenarios. Scenarios FN5 and REJ5 are annotated as non-aggressive hate speech targeted at a generic group and seem less hateful than the other posts, and the neutral post in REJ7 contains an offensive slur. Given the nature of these social media posts used in these scenarios, it might explain the larger disagreement between participants in annotating it as hateful/non-hateful for these scenarios.

We simulated the TP, TN, FP, FN, and rejection scenarios by asking the second and third questions where the participant had to provide a response value using either the ME or the 100-level scale to express their agreement or disagreement with SocialNet's decision. We analyze the resulting response values by looking at three aspects. First, we analyze the value ratios from the survey that uses the ME scale in section 6.1.1. Then we discuss whether the ME technique passes the reliability and validity analyses in sections 6.1.2 and 6.1.3. Finally, we analyze the results of the demographic analysis in section 6.1.4

6.1.1 Value ratios

Regarding the value ratios, most results align with our hypotheses. The resulting values of both the ME and the 100-level scale follow the same order ($V_{fn} < V_{fp} < V_r < V_{tp} < V_{tn}$). We noticed that participants disagreed the most with scenarios FN3 and FN7. According to the annotations given by Basile et al. (2019), both scenarios are hate speech targeted at an individual and contain aggressive speech. The results of FN3 and FN7 might suggest that participants are more likely to disagree with FN predictions for aggressive hate speech targeted at individuals.

As expected, we found that participants disagree with the FP, FN, and rejected predictions, that the value of an FN is lower than an FP, and that the average value of an FP and an FN is lower than the rejection value. The results show that participants find a hateful post that is not detected worse than a non-hateful post detected as hate speech. This finding is in line with our hypothesis that tolerating hate speech (FN predictions) harms social media users more than forbidding neutral speech (FP predictions). The value of rejection is the closest to 0 (neutral) because, according to our formulations in the survey, rejected predictions only reduce the value by the human moderation effort and do not cause much benefit or harm since the human moderator needs to handle the prediction within 24 hours.

However, two things were somewhat surprising. First, participants appreciated correct predictions more than incorrect predictions since participants gave higher absolute values to TP and TN scenarios than FP and FN scenarios. We expect participants to give lower absolute response values to correct predictions since it is expected from the automatic detection algorithms to produce correct predictions. However, we look at this from a computer science perspective, where we want to prevent incorrect predictions, whereas the participants might think producing correct predictions is more critical. Second, we were surprised that the TN value was greater than TP, while we expected the opposite to hold. One possible reason could be that people disagree more on what is considered hateful among the TP scenarios, as seen in figure 5.1. This observation is in line with the findings of Ross et al. (2017),

as the authors found low agreement among participants regarding labelling social media posts as hate speech.

6.1.2 Reliability analysis

According to the Krippendorff's alpha values (α), the results of the ME scale are reliable, indicating that the ME technique is suitable for estimating the value ratios. Contrary to our hypothesis, the results indicated that the 100-level scale is less reliable than the ME scale. We would expect many participants to give response values of -100, 100, or 0 as the 100-level scale is bounded, and, therefore, we would expect higher alpha values for the 100-level scale compared to the ME scale.

In general, the results also showed low alpha values when we computed it for each group of scenarios with the same type (TP, TN, FP, FN, and rejection). Users tend to agree more on incorrect predictions than on correct predictions, indicating that participants agree more on the harm caused by incorrect predictions. We can explain the low reliability values by looking at the calculation of Krippendorff's alpha. In this calculation, we measure the difference between the expected difference and the observed difference. When we consider the response values to all scenarios, the values tend to follow the same trend; positive values for correct predictions and negative values for incorrect and rejected predictions. When we consider the response values to the scenarios of the same type, e.g. all TP scenarios, the values seem more randomly distributed as each participant uses a different positive response value to the TP scenarios. Therefore, when considering all scenarios, the observed difference between the response values is closer to the difference expected by chance, resulting in higher alpha values.

6.1.3 Validity analysis

The cross-modality validation between the ME and the 100-level scales showed that the response values to both scales are highly correlated, indicating that we validated the ME technique for measuring people's opinions about different hate speech detection scenarios. The S-shaped curve in figure 5.3 is because for two reasons. First, the magnitude estimates are skewed towards 0 because of the normalization procedure. Second, the 100-level scores are skewed towards the upper and lower bounds of 100 and -100 as the participants are more likely to assign the highest or lowest value.

6.1.4 Demographic analysis

We analyzed several demographic features (sex, student, continent, nationality, language, and ethnicity) to see if significant differences exist between groups of participants in the response values to all scenarios.

For all scenarios, we found no differences between men and women. This finding is in line with the work of Gold and Zesch (2018), as the authors did not find any differences between men and women and how they perceive hate.

For the remaining five features, we found significant differences between groups of participants for only a small number of scenarios. Furthermore, for the scenarios and features with more than two groups (nationality, language, and ethnicity) where we found significant differences, we often did not find any significant pairwise differences between the groups. These results indicate that for our dataset, people with different demographic characteristics tend to give the same judgements

to different hate speech detection scenarios. Nevertheless, the results show that people with different nationalities, languages, and ethnicities are more likely to differ in their opinions about hate speech detection scenarios than people of different sex or student status.

We found the most group differences for scenarios FP7 and REJ4 (both containing non-hateful posts) among all features. It is unclear why FP7 had so many significant differences, as the post is neutral and not about any sensitive topic. However, the social media post used in REJ4 is about refugees, which can be a politically sensitive topic. People with different demographic characteristics, such as continent, language, nationality, or ethnicity, could have different opinions about this topic. There were few pairwise group differences for both scenarios and the features of nationality and language. However, we observed differences for two of the three pairwise combinations for the ethnicity feature and both scenarios. Nevertheless, given these results, there is not enough evidence to explain why scenarios such as FP7 and REJ4 cause more group differences than other scenarios.

Also, we found that hateful social media posts are more likely to cause group differences than non-hateful posts, as we have more scenarios with group differences that contain hateful posts (10 in total) than non-hateful posts (5 in total).

We observed the most pairwise significant differences for scenario FN5. Scenario FN5 contains a hateful social media post about building the wall across the border between the United States and Mexico. There are five posts about building the wall, both hateful and non-hateful. For four out of the five posts, we found at least one feature with significant differences between the groups of participants, suggesting that group differences depend on the topic of the social media post.

6.2 Value-sensitive rejection

We analyze three aspects of our value-sensitive rejector. First, we analyze how the rejector behaves when applied to different hate speech detection models and datasets (the *seen* and *unseen* datasets). The *seen* dataset is a test set sampled from the same dataset as the training set. In contrast, the *unseen* dataset is a test set sampled from a completely different dataset to simulate how the models perform on new and unfamiliar data. Second, we study whether value-sensitive rejection of ML predictions can be beneficial for hate speech detection. Finally, we compare our value-sensitive metric to machine metrics such as accuracy.

We observed that the three hate speech detection models are not well-calibrated, meaning many high-confident incorrect and low-confident correct predictions exist. Therefore, when we apply a rejection threshold, we have the problem of accepting many incorrect predictions or rejecting many correct predictions. Nevertheless, we observed that the models are more confident in the correct predictions than the incorrect predictions, making the value-sensitive rejector still useful.

The results of our value-sensitive metric were very similar for all three models and both datasets. When we consider all value ratios, accepting all predictions seems the most valuable for both the *seen* and *unseen* data. This result is not surprising as the absolute magnitudes of TP and TN are greater than the absolute magnitudes of FP and FN, and there are more TP and TN predictions than FP and FN predictions. Therefore, the gains of accepting all correct predictions outweigh the costs of accepting all incorrect predictions for all models and datasets.

We believe it is more critical to focus on punishing incorrect predictions, as we want to minimize harm to social media users. Therefore, in the second part of the

experiments with the value-sensitive rejector, we no longer focused on rewarding correct predictions, implying $V_{tp} = 0$ and $V_{tn} = 0$. As a result and according to the formulation of the value-sensitive metric (formula 3.5), accepted correct predictions increase the total value by the value of rejection (V_r) and correct predictions that are rejected decrease the total value by the value of rejection. For the *seen* data, the results of the optimal rejection threshold show that by not rewarding correct predictions, a significant fraction of the predictions can be accepted from all three models and a smaller fraction rejected. All three models with the optimal rejection threshold are valuable for the *unseen* data, but very few predictions can be accepted, and the majority are rejected. The high optimal rejection thresholds for the *unseen* data also demonstrate that hate speech detection models are susceptible to bias, in line with the findings of related studies by Arango et al. (2019) and Gröndahl et al. (2018). When we accept all predictions, all three models are valuable for the *seen* data but invaluable for the *unseen* data, putting the viability of all models into question. Therefore, the results show that our value-sensitive rejector can benefit hate speech detection and help us determine when to rely on the ML models.

Finally, we compared the results of our value-sensitive metric with machine metrics like accuracy. If we accept all predictions, we find that both metrics indicate that the DistilBERT model performed the best. However, when we consider not rewarding correct predictions for *unseen* data, both metrics return different results. According to our value-sensitive metric, the CNN model is the best, while accuracy indicates that either the LR or DistilBERT model is the best. We think that the CNN model has a higher total value as it produces fewer FN predictions (which are costly) than the other two models.

We see some interesting things when we compare the value-sensitive metric for the optimal rejection thresholds with the accuracy metric. For most configurations, both metrics return the same results, namely that the DistilBERT model is the best. However, when considering not rewarding correct predictions for the *unseen* data, we see that the LR model performs the best and gets a slightly higher total value than the DistilBERT model for the optimal rejection threshold. What makes this finding interesting is that while the accuracies of the original models are the same, we would expect that the DistilBERT has a higher total value because the DistilBERT model has a higher accuracy of the accepted predictions and a lower rejection rate. One explanation might be that we found that the LR model achieves a higher total value since it rejects more FN predictions and accepts fewer FN predictions than the DistilBERT model for the optimal rejection threshold.

6.3 Implications

Related studies recognized machine metrics' shortcomings, such as accuracy (Casati et al., 2021; Olteanu et al., 2017; Röttger et al., 2020). While Röttger et al. (2020) focused on evaluating hate speech detection models by presenting a suite of automated tests, we focused on improving existing hate speech detection models by adopting a reject option. Olteanu et al. (2017) claim that we need more human-centred metrics that take the perceived cost of incorrect decisions into account instead of using abstract metrics such as precision. They state that this perceived cost should depend on the context of the specific problem and the type of incorrect decisions (Olteanu et al., 2017). Our work aligns with theirs as we presented a value-sensitive metric that considers human value and conducted a survey study to retrieve the perceived value of social media users for TP, TN, FP, FN, and rejected

predictions in hate speech detection. Casati et al. (2021) promoted the use of value-sensitive metrics. They recognized the limits of machine metrics such as accuracy, as two models could have the same accuracy, but one model could be more valuable than the other (Casati et al., 2021). Our experiments demonstrated this difference as we found that the best model according to our value-sensitive metric could be different from the best model according to the accuracy metric. We believe these findings can benefit industry and research, as many domains exist where tasks cannot be fully automated and where human-AI solutions, such as our value-sensitive rejector, can increase the utility of ML models.

Regarding the survey study, we got several interesting findings. Social media users appreciate correct hate speech predictions more than they detest incorrect predictions. However, in terms of inter-rater reliability, they agree more on recognizing the harm caused by incorrect predictions than the gain from correct predictions. Overall, the inter-rater reliability values of all scenarios of the same type were low. This observation is in line with the findings of Ross et al. (2017), where the authors also found low Krippendorff's alpha values when asking participants to annotate hate speech. We did not find many significant differences between groups of different demographical characteristics. Like Gold and Zesch (2018), we did not find any differences between men and women and how they perceive hate. However, we found more differences when looking at other demographical features, such as nationality, language, or ethnicity, implying that these features are more likely to cause group differences.

Regarding the hate speech detection models, we found that BERT models are indeed promising for hate speech detection, given the recent popularity of BERT models for NLP applications (Alatawi et al., 2021; Edwards, 2021). The results with all three hate speech detection models also demonstrated the impact of dataset bias since we found significant performance drops in terms of both the value of our value-sensitive metric and the accuracy metric. The experiments with the *unseen* data resulted in lower total values and accuracies compared to the *seen* data, indicating that hate speech datasets are biased. Once we train hate speech detection models on such biased datasets, the models also become biased. Our results fit the findings of previous studies where the authors found a significant performance drop when they trained models on one dataset and evaluated them on another (Arango et al., 2019; Gröndahl et al., 2018).

Regarding the methodology, we believe the ME technique is interesting for social science-related problems where the goal is to retrieve human-perceived value ratios. We showed how we could use the value ratios in a value-sensitive metric that measures the total human-perceived value of ML models with a reject option. We further demonstrated how we could create a human-AI solution for hate speech detection by using the value-sensitive metric to calculate the optimal rejection threshold. We used the optimal rejection threshold to determine when we could trust machine predictions and when we needed to pass machine predictions to a human moderator.

6.4 Limitations

In this section, we list the limitations of the survey study and the value-sensitive rejector.

Regarding the survey study, we had a limited sample size of 68 participants per scale due to a constrained budget. We expect more reliable results when experimenting with larger sample sizes.

Second, we limited the number of scenarios to eight per type, each including either a hateful or non-hateful post depending on the scenario type. We expect the results to be more reliable if the experiment had been repeated several times with additional sets of social media posts for multiple groups of participants. Nevertheless, we believe the results are still reliable since we performed a content analysis procedure for selecting the most representative social media posts for our experiment.

Third, dealing with hateful content on social media platforms remains controversial, even for governmental institutions and social media companies. We believed the results would differ when we used different descriptions in the scenarios. Initially, we explained in the pilot survey that SocialNet removes hateful posts. After gathering the results, we noticed that participants assigned larger absolute values to the TP and TN scenarios than to the FP and FN scenarios. Therefore, we decided to update the descriptions to rank posts lower instead of removing posts and explained that it is expected from detection systems to produce correct predictions and that incorrect predictions might cause harm to social media users. After updating the descriptions, we did not notice any difference as participants still assigned larger absolute values to TP and TN scenarios. Nevertheless, we still believe that using different descriptions would give different results as we think that people have different opinions about how we should deal with detected hate speech.

Finally, we should point out the limitations of the demographic analysis. We did not apply any demographical constraints when gathering participants for the survey study. As a result, the demographical characteristics of the participants were entirely random, and the sample sizes were relatively small. For example, most participants in our experiment lived in South Africa or Poland, and we only had five participants from Spain. Although the sample sizes were large enough for the statistical tests, they were not large enough to represent the populations of entire countries. We did not find enough evidence that people with different demographic characteristics have different opinions about hate speech detection scenarios. However, if we repeated the experiment with larger sample sizes, we expect to find more group differences for some demographical features, such as nationality. At the same time, we also believe that for the features where we did find significant differences between demographic groups, some of them might have happened by chance. Either because participants did not understand the scenario, their lack of English, or because they rushed through the survey.

Regarding our value-sensitive rejector, we believe our approach has several limitations.

First, the rejection threshold is calculated empirically and depends highly on the choice of the test dataset. As we have seen in our experiments, we retrieved different optimal rejection thresholds for different test datasets (the *seen* and *unseen* datasets). Factors such as *sample retrieval* or *sample annotation* bias (refer to section 2.1) explain why we got different results for the *seen* and *unseen* datasets. Therefore, when using the value-sensitive rejector, it is essential to use a test set that is as similar to real-world data as possible.

Second, we think using well-calibrated models in the value-sensitive rejector is best. Although calibration methods such as temperature scaling can improve existing classification models, these techniques are limited as we still observed many high confident errors.

6.5 Recommendations

We believe that future research on social-science-related problems, such as hate speech detection, should focus more on creating human-AI collaboration solutions and that these solutions should take human value into account. We noticed that most research in hate speech detection focuses solely on creating fully-automated classification models with accuracy as its optimization target. We think solutions such as ML with rejection are promising to increase the utility of classification models for tasks that cannot be fully automated, such as hate speech detection. Furthermore, we showed the limitations of machine metrics, such as accuracy, as they do not consider the context of the problem. We found through our survey study that different types of machine errors have different costs according to the social media users. Therefore, we think that future work should also focus more on developing value-sensitive or human-centred metrics.

Given the limitations of our survey study, we suggest repeating our experiment with larger sample sizes to increase the reliability of the results. Also, we think it would be interesting to study which factors influence the user perception of hate speech detection scenarios. We believe that user perception depends on many factors, such as the scenario's description, the post's topic, whether the post is offensive or not, and the post's target(s). Finally, we think it would be interesting for future research to study the effects of demographical characteristics in more detail by repeating the experiment for different demographic groups with larger sample sizes.

Given the limitations of our value-sensitive rejector, it would be interesting to create a hybrid solution of our value-sensitive rejector and (un)known unknown detection techniques. By (un)known unknowns, we mean the low confident correct and high confident incorrect predictions. If the underlying classification model of our value-sensitive rejector is not well-calibrated, then we end up with many (un)known unknowns. We suggest future work to combine (un)known unknown detection techniques with our value-sensitive rejector so that less correct and more incorrect predictions are rejected and that more correct and less incorrect predictions are accepted. Finally, as the optimal rejection threshold is calculated empirically on a hate speech dataset, we should prevent ourselves from using biased datasets. We used datasets that were collected using specific keywords and annotated by only three annotators. Therefore, we suggest using datasets where sample retrieval and sample annotation bias are prevented as much as possible, for example, by collecting only the most representative data samples or by annotating the data by a large group of annotators with diverse demographic characteristics.

Chapter 7

Conclusion

This research aimed to tackle the problems of automatic and manual proactive moderation of hate speech on social media platforms. We presented a human-AI solution for hate speech detection where we reject machine learning (ML) predictions in a value-sensitive manner. In the first half of this project, we formulated a value-sensitive metric for measuring the total value of an ML model with a reject option based on the implications of true positive (TP), true negative (TN), false positive (FP), false negative (FN), and rejected predictions. We used the value-sensitive metric to determine the optimal confidence threshold for which the model achieves the maximum total value. In practice, we accept all ML predictions with confidence values above the optimal threshold and reject all below the threshold so that the human moderator makes the final judgement. In the second part, we designed a survey study to determine the value ratios between TP, TN, FP, FN, and rejected predictions in the context of hate speech detection from the perspective of social media users. We proposed using the magnitude estimation (ME) scale for measuring user perception of different hate speech detection scenarios.

The survey study uncovered several findings. We showed that ME is a reliable technique for measuring the value ratios. We validated the results by showing the correlation with the results from a separate survey study using a 100-level scale. We found that participants mostly appreciate the correct predictions while strongly agreeing with the harm of incorrect predictions. The demographic analysis showed that participants with different demographical characteristics do not differ in their perception of most hate speech detection scenarios. However, these results are solely indicative due to the small sample sizes of the demographic groups in our data.

To demonstrate the utility of our value-sensitive rejector, we experimented with three state-of-the-art hate speech detection models on two real-world hate speech datasets. We got several interesting findings when focusing on minimizing harm. The results show the utility of our value-sensitive rejector in guiding the decision of accepting or rejecting machine learning predictions. The results show that the value-sensitive rejector maximizes the utility of hate speech detection models when optimizing value for data familiar to the model's training data, as only a small fraction of the predictions were rejected. In line with the findings from related research, we found that hate speech detection models are susceptible to bias, resulting in a large fraction of the predictions being rejected when applying the value-sensitive rejector to unfamiliar data.

Appendix A

Survey

This appendix contains all the presentation material of the survey: scenarios, consent, explanation texts, and some examples of scenarios.

A.1 Scenarios

No.	Text	H	T	A
TP1	CBS: Driver an Illegal Alien was drunk, high at time of fatal Donna crash that killed 6 year old girl. Yet another victim whose life was stolen by an illegal alien.When will you end this?#BuildTheWall #NoAmnesty #NoDACA #DeportThemAll	Y	G	Y
TP2	What do you call a woman who has lost 95% of her intelligence? Divorced	Y	G	N
TP3	first goes to bitch ass janelle. FUCK YOU MESSY MF ASS BITCH i literally was on a call with her and her dumb ass twisted my words and ran to tattle on me to ari. like it legit made her pussy wet to go and talk shit abt me so fuck u ill come beat ur ass thats why u stole a listers	Y	I	Y
TP4	Soccer players go down more than my bitch on my dick	Y	I	N
TP5	#BuildThatWallNow I do not want those vile thugs in our country! #EndChainMigration #EndSanctuaryCities #EndVisaLottery #AngelFamlies	Y	G	Y
TP6	Crimes committed by ILLEGAL immigrants should be public information everywhere like in Texas. More illegal aliens were arrested for homicide in the past 7 years in a single state than deaths by mass shooting in the other 49 states combined! #SendThemBack #BuildTheWall	Y	G	N
TP7	You smoke too much stfu pussy	Y	I	Y
TP8	Your skank ass disgusts me so much	Y	I	N

TABLE A.1: All true positive scenarios used for the survey. The texts are from the Basile et al. (2019) dataset. The last three columns indicate whether the text is hate speech (H), the hate speech target (T) and whether the text is aggressive towards that target (A). 'Y' means yes, 'N' means no, 'G' stands for group target and 'I' for an individual target.

No.	Text	H	T	A
TN1	How many immigrant kids were reunited with their parent today?	N	-	-
TN2	I like to tell people that I drink every night cause I can't sleep but at this point it's just a blatant lie cause I've been tired for years	N	-	-
TN3	It's what I do, it's who I am	N	-	-
TN4	Just walked past this women and she goes 'Hello you cunt' I'm crying :(:(	N	-	-
TN5	#MeToo Not all men, far from it, sexually abuse women.However, all women have experienced sexual harasment.So not all men, but all women.	N	-	-
TN6	I never smile and I hate to brag but I can say 'cheese' without smiling too	N	-	-
TN7	Our Table at the Museum of Jewish Heritage - A Living Memorial to the Holocaust For the Immigrant Arts & Women's Empowerment Summit	N	-	-
TN8	You can't be pro -sexual liberation (for women especially) and be anti -hoe by calling people you don't like slut, hoe, whore	N	-	-

TABLE A.2: All true negative scenarios used for the survey. The texts are from the Basile et al. (2019) dataset. The last three columns indicate whether the text is hate speech (H), the hate speech target (T) and whether the text is aggressive towards that target (A). 'Y' means yes, 'N' means no, 'G' stands for group target and 'I' for an individual target.

No.	Text	H	T	A
FP1	I was never properly taught how to put on makeup. The only looks I have are fancy whore or daytime slut.	N	-	-
FP2	omg and pussy boys voicemail wishing me a happy birthday was the absolute best I appreciate you so much, thank you pussy boy <3	N	-	-
FP3	Ontario concerns were not listened at immigration meeting in Ottawa For illegal crossing	N	-	-
FP4	Rape is DISGUSTING. No woman or man deserves to go through something as traumatic as that. Nothing can ever justify rape. Rape is a choice.	N	-	-
FP5	Ann Coulter calls Cory Booker a 'hysterical woman'	N	-	-
FP6	How far can you go down the rabbit hole ?	N	-	-
FP7	I don't run from my problems, I laugh at them and act like they don't exists like a normal human being	N	-	-
FP8	Is there historical precedent for a state killing its own people, as a consequence of which many of them become refugees, then, without providing any guarantees or safety, they ask for people to return and the neighboring countries try to push people back into the dictatorship?	N	-	-

TABLE A.3: All false positive scenarios used for the survey. The texts are from the Basile et al. (2019) dataset. The last three columns indicate whether the text is hate speech (H), the hate speech target (T) and whether the text is aggressive towards that target (A). 'Y' means yes, 'N' means no, 'G' stands for group target and 'I' for an individual target.

No.	Text	H	T	A
FN1	Why do women have periods? Because they deserve them	Y	G	Y
FN2	Fuck hoes a bitch ain't nothin women be trippin bruh	Y	G	N
FN3	I want this stupid bitch out of my head I've never even liked her lmao I just basically want to rape her	Y	I	Y
FN4	Hoe asss bitch what the fuck	Y	I	N
FN5	We Need a WALL #BuildThatWall	Y	G	Y
FN6	Don't you hate how girls just Whore around because they never got daddy's attention ?? And so they have to interfere in relationships.. hoe	Y	G	N
FN7	You mother fucking piece of trash stupid fucking idiot hoe i hope you fucking die in hell son of a father fucking fuck bitch whore you fucki	Y	I	Y
FN8	It's like it's always the ones YOU CARE FOR THE LEAST that be talking so much shit like bitch I'm not even paying attention to you lil whore	Y	I	N

TABLE A.4: All false negative scenarios used for the survey. The texts are from the Basile et al. (2019) dataset. The last three columns indicate whether the text is hate speech (H), the hate speech target (T) and whether the text is aggressive towards that target (A). 'Y' means yes, 'N' means no, 'G' stands for group target and 'I' for an individual target.

No.	Text	H	T	A
REJ1	niggas think bitches supposed to suck their lil ass dick and not give a bitch head too, like that's not how it works whore !	Y	G	Y
REJ2	FUCKING FUCK SHIT PISS MOTHERFUCKING BITCH TIT CUNT FUCKING SLUT	Y	I	N
REJ3	Ever been so hungover that your stomach feels like it's eating itself	N	-	-
REJ4	Lifesaving aid on Cabeza is a drop of water in the desert: precious but not nearly enough. We must stop the forces that force people to cross in the West Desert: #WelcomeRefugees, #SafePassage for ALL migrants, stop the economic & political violence the US wages on Latin America	N	-	-
REJ5	Only men have a dick they can insert without consent, and women doing other shit. Not that women can't rape & sexually harass either.	Y	G	N
REJ6	Bitch shut the fuck up and dont talk about Afghanistan	Y	I	Y
REJ7	Pokemon Gooooooo fuck yourself	N	-	-
REJ8	Jesus Sex Trafficking is grim. A lot of these women were damaged by something happening to them and that's where the pimp took advantage	N	-	-

TABLE A.5: All rejection scenarios used for the survey. The texts are from the Basile et al. (2019) dataset. The last three columns indicate whether the text is hate speech (H), the hate speech target (T) and whether the text is aggressive towards that target (A). 'Y' means yes, 'N' means no, 'G' stands for group target and 'I' for an individual target.

A.2 Consent

You are being invited to participate in a research study titled "Costs of predictions in hate speech detection". This study is being done by Philippe Lammerts from the TU Delft.

The purpose of this research study is to find out what social media users think of different scenarios of hate speech detection on social media. It will take you approximately 22 minutes to complete. These scenarios consist of two things. First, we show a specific social media post that can be either hateful or not hateful. You need to indicate if you feel that the post is hateful or not. Second, we explain how the social media platform dealt with this post. You need to indicate whether you agree/disagree/are neutral about the platform's decision. The results of the survey will be used in my thesis.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by making this survey completely anonymous. Therefore, please do not provide any personal information anywhere. The anonymous results might be shared publicly in the future.

Your participation in this study is entirely voluntary, and you can withdraw at any time.

Warning: some of the scenarios used in this experiment contain harmful and offensive content that may make some people feel uncomfortable.

Feel free to contact me with any questions or feedback you might have:
p.m.lammerts@student.tudelft.nl

A.3 Introduction

A.3.1 ME scale

- You will be presented with a series of different scenarios.
- For each scenario, you need to answer two questions.
- We will explain the exact instructions later.
- But first, we will let you familiarize yourself with a scale called Magnitude Estimation.

A.3.2 100-level scale

- You will be presented with a series of different scenarios.
- For each scenario, you need to answer two questions.
- We will explain the exact instructions in the next page.

A.3.3 Introduction

You will be presented with a series of different scenarios.

- Each scenario describes a situation of a social media user who wants to post a specific message on a fictional social media platform we now call SocialNet.
- These posts can be neutral or contain hateful content.
- SocialNet uses automated detection systems for detecting hate speech.
- When doing the study, you should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Each scenario describes one of the following situations for a specific social media post:

1. **You are a user of the SocialNet platform and have not seen this post on your main feed because SocialNet's automated detection system is confident that it is hateful.**
    - You can still find this post when you scroll down your feed since SocialNet ranks hateful posts lower.
    - If the post is not hateful after all, then the detection system was incorrect. This neutral post is now ranked lower on people's feeds with the consequence that the post cannot easily reach the author's followers.
    - If the post is indeed hateful, then the detection system was correct.
2. **You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system is confident that it is not hateful.**
    - This post remains visible on other people's main feeds as well.
    - If the post is hateful after all, then the detection system was incorrect. This hateful post is now visible on people's main feeds with the consequence that they can get harmed.
    - If the post is indeed not hateful, then the detection system was correct.
3. **You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system was not confident enough in whether it was hateful or not.**
    - An internal human moderator at SocialNet needs to look at it within at most 24 hours.
    - Meanwhile, the post remains visible on people's main feeds.

A.4 Scale explanations

A.4.1 ME scale

The following text is based on the survey setup from Moskowitz (1977).

For each scenario, you need to answer two questions:

1. First, you need to indicate whether you feel that this post is hateful or not hateful.
2. Second, your task is to tell how you feel about SocialNet's decision.
    - If you feel neutral about SocialNet's decision, this value will be equal to 0.
    - If you (dis)agree with the decision from SocialNet, you need to assign any number that is greater or equal to 0 that reflects how much you (dis)agree with the decision.
    - Assign any number that seems appropriate to you.
    - A large number means you (dis)agree a lot, while a small number means you (dis)agree a little.
    - If you (dis)agree twice as much with the current decision as with the previous one, you need to assign a number that is twice as large as the previous number.
    - Or, if you (dis)agree half as much with the current decision as with the previous one, you need to assign a number that is half as large as the previous number.
    - You can use any number or decimal you want, but make each assignment proportional to your subjective impression.
    - Please try to not only use appropriate numbers but also avoid restricting your choice of numbers from 1 to 10.
    - Try to remember which values you used before and to compare succeeding scenarios with each other.

Don't worry, we will provide the same explanations in the questions as well.

A.4.2 100-level scale

For each scenario, you need to answer two questions:

1. First, you need to indicate whether you feel that this post is hateful or not hateful.
2. Second, your task is to tell how you feel about SocialNet's decision.
    - If you feel neutral about SocialNet's decision, this value will be equal to 0.
    - If you (dis)agree with the decision, you need to indicate how much you (dis)agree by assigning any number between 1 and 100.
    - A large number means you (dis)agree with it a lot, while a small number means you (dis)agree with it a little.
    - Try to make each number match the intensity as you perceive it.

Don't worry, we will provide the same explanations in the questions as well.

A.5 Training phase ME

"As a warm-up task, to familiarize you with magnitude estimation, you will be shown a sequence of five lines, one at a time.

- For each line, enter a number into the text box below the displayed line. This number should reflect your perception of the length of the line. You may use any numbers that seem appropriate to youwhole numbers or decimals. However, you may not use negative numbers or zero.
- For each subsequent line, enter a number that reflects your perception of its length, relative to the previous line. For example, if you feel that the current line is twice as long as the previous, then you should assign a number that is twice as large as the number you used previously.

Don't worry about running out of numbersthere will always be a larger number than the largest you use, and a smaller number than the smallest you use. Note: The magnitude estimation scores are **not** intended to be an estimate of the length in any particular measurement units, such as centimeters." (Maddalena et al., 2017)

A.6 Examples

A.6.1 FN scenario with ME scale

WARNING: the example used in this section contain content that may make some people feel uncomfortable.

Suppose we have the following post:
"I want this stupid bitch out of my head I've never even liked her lmao I just basically want to rape her" (from Basile et al. (2019))

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system is confident that it is not hateful.

- This post remains visible on other people's main feeds as well.
- If the post is hateful after all, then the detection system was incorrect. This hateful post is now visible on people's main feeds with the consequence that they can get harmed.
- If the post is indeed not hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.
Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number.
Show ME input field that allows all positive values greater than 0.

A.6.2 FP scenario with 100-level scale

Suppose we have the following post:
"Ontario concerns were not listened at immigration meeting in Ottawa For illegal crossing" (from Basile et al. (2019))

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and have not seen this post on your main feed because SocialNet's automated detection system is confident that it is hateful.

- You can still find this post when you scroll down your feed since SocialNet ranks hateful posts lower.
- If the post is not hateful after all, then the detection system was incorrect. This neutral post is now ranked lower on people's feeds with the consequence that the post cannot easily reach the author's followers.
- If the post is indeed hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.
Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number from 1 to 100. If you feel neutral about SocialNet's decision, select neutral in the field above.
Show a numerical slider with values between 1 and 100.

A.6.3 Rejection scenario with 100-level scale

Suppose we have the following post:
"Ever been so hungover that your stomach feels like it's eating itself" (from Basile et al. (2019))

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system was not confident enough in whether it was hateful or not.

- An internal human moderator at SocialNet needs to look at it within at most 24 hours.
- Meanwhile, the post remains visible on people's main feeds.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.
Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number.
Show a numerical slider with values between 1 and 100.

Appendix B

Results

This appendix contains the remaining demographic analysis of the survey results from section 5.1.4 and the experiments with the value-sensitive rejector from section 5.2.

B.1 Demographic analysis

This section contains some additional tables about the demographic analysis. Table B.1 shows the group differences for the aggregated scenario types (TP, TN, FP, FN, and REJ). Tables B.2, B.3, and B.4 show the pairwise group differences for the nationality, language, and ethnicity features, respectively.

	Two groups			More than two groups		
	Sex	**Student**	**Continent**	**Nationality**	**Language**	**Ethnicity**
TP	0.302	**0.032**	0.286	0.218	0.109	0.242
TN	0.726	0.379	0.204	0.190	0.216	0.281
FP	0.699	0.933	0.073	**0.020**	**0.040**	**0.037**
FN	0.961	0.150	0.847	0.478	0.438	0.584
REJ	0.835	0.625	0.496	0.271	0.103	0.068

TABLE B.1: **Aggregated**: an overview of the statistical differences between different groups of participants for various demographic characteristics for each aggregated scenario type in the ME survey. Each cell contains the p value of either the Mann-Whitney U test for two groups or the Kruskal-Wallis test for more than two groups. The grey cells with bold text indicate significant statistical differences between the groups for that feature and scenario type.

	South Africa	Poland	Portugal	Spain
South Africa	1.000			
Poland	0.077	1.000		
Portugal	0.077	**0.009**	1.000	
Spain	0.119	0.083	0.613	1.000

(A) TP6

	South Africa	Poland	Portugal	Spain
South Africa	1.000			
Poland	0.755	1.000		
Portugal	0.261	0.261	1.000	
Spain	**0.026**	**0.038**	**0.050**	1.000

(B) FP2

	South Africa	Poland	Portugal	Spain
South Africa	1.000			
Poland	0.342	1.000		
Portugal	0.304	1.000	1.000	
Spain	**0.043**	0.304	0.220	1.000

(C) FP7

	South Africa	Poland	Portugal	Spain
South Africa	1.000			
Poland	**0.034**	1.000		
Portugal	0.150	**0.011**	1.000	
Spain	**0.045**	**0.011**	0.679	1.000

(D) FN5

	South Africa	Poland	Portugal	Spain
South Africa	1.000			
Poland	0.095	1.000		
Portugal	0.622	0.095	1.000	
Spain	0.104	0.095	0.104	1.000

(E) FN6

	South Africa	Poland	Portugal	Spain
South Africa	1.000			
Poland	0.088	1.000		
Portugal	0.227	0.088	1.000	
Spain	1.000	0.227	0.388	1.000

(F) REJ1

	South Africa	Poland	Portugal	Spanish
South Africa	1.000			
Poland	0.098	1.000		
Portugal	0.098	1.000	1.000	
Spain	0.098	0.422	0.422	1.000

(G) REJ4

TABLE B.2: **Nationality**: an overview of all pairwise Mann-Whitney U tests between the different nationalities for all scenarios where we found significant differences between all nationalities using the Kruskal-Wallis test. Each cell contains the p value of the Mann-Whitney U test between two groups of different nationalities. We corrected all p values with the Benjamini-Hochberg procedure. The grey cells with bold text indicate significant statistical differences between the two nationalities.

	English	Polish	Portugese	Spanish
English	1.000			
Polish	0.561	1.000		
Portugese	0.352	0.283	1.000	
Spanish	0.176	0.258	0.142	1.000

(A) TP3

	English	Polish	Portugese	Spanish
English	1.000			
Polish	0.089	1.000		
Portugese	0.119	**0.007**	1.000	
Spanish	0.522	0.089	1.000	1.000

(B) TP6

	English	Polish	Portugese	Spanish
English	1.000			
Polish	0.321	1.000		
Portugese	0.321	1.000	1.000	
Spanish	**0.019**	0.444	0.321	1.000

(C) FP7

	English	Polish	Portugese	Spanish
English	1.000			
Polish	0.070	1.000		
Portugese	0.209	**0.011**	1.000	
Spanish	0.647	0.164	0.838	1.000

(D) FN5

	English	Polish	Portugese	Spanish
English	1.000			
Polish	0.895	1.000		
Portugese	0.439	0.721	1.000	
Spanish	**0.049**	0.309	0.548	1.000

(E) FN7

	English	Polish	Portugese	Spanish
English	1.000			
Polish	0.076	1.000		
Portugese	0.387	0.076	1.000	
Spanish	0.149	0.711	0.096	1.000

(F) REJ1

	English	Polish	Portugese	Spanish
English	1.000			
Polish	0.063	1.000		
Portugese	0.063	1.000	1.000	
Spanish	0.599	1.000	1.000	1.000

(G) REJ4

	English	Polish	Portugese	Spanish
English	1.000			
Polish	1.000	1.000		
Portugese	0.489	0.452	1.000	
Spanish	0.105	0.152	0.105	1.000

(H) REJ5

TABLE B.3: **Language**: an overview of all pairwise Mann-Whitney U tests between the different spoken languages for all scenarios where we found significant differences between all spoken languages using the Kruskal-Wallis test. Each cell contains the p value of the Mann-Whitney U test between two groups of languages. We corrected all p values with the Benjamini-Hochberg procedure. The grey cells with bold text indicate significant statistical differences between the two languages.

	White	Mixed	Black
White	1.000		
Mixed	0.552	1.000	
Black	**0.028**	**0.028**	1.000

(A) FP7

	White	Mixed	Black
White	1.000		
Mixed	0.776	1.000	
Black	**0.002**	**0.776**	1.000

(B) REJ4

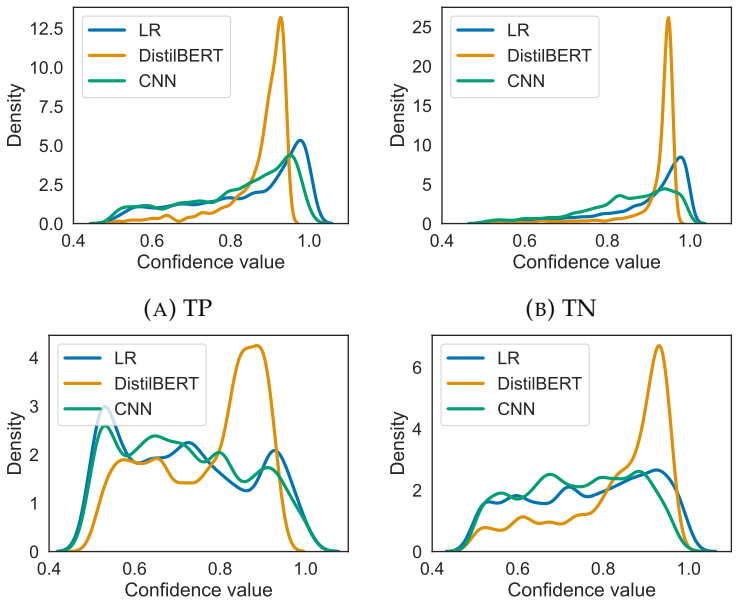
	White	Mixed	Black
White	1.000		
Mixed	**0.016**	1.000	
Black	1.000	0.112	1.000

(C) REJ7

TABLE B.4: **Ethnicity**: an overview of all pairwise Mann-Whitney U tests between the different ethnicities for all scenarios where we found significant differences between all ethnicities using the Kruskal-Wallis test. Each cell contains the p value of the Mann-Whitney U test between two groups of ethnicities. We corrected all p values with the Benjamini-Hochberg procedure. The grey cells with bold text indicate significant statistical differences between the two ethnicities.

B.2 Probability density functions

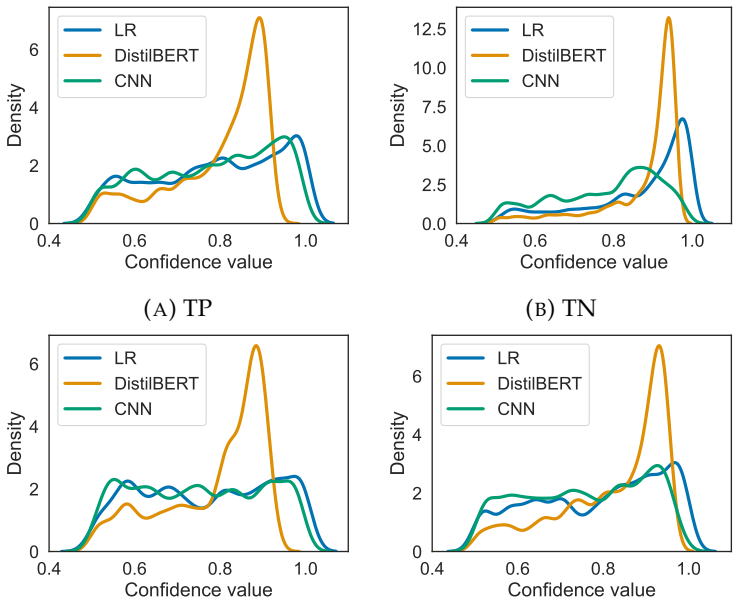
This section contains the probability density functions of the predictions of all models to both the *seen* (figure B.1) and the *unseen* (figure B.2) datasets. They help us understand the value-sensitive metric results in figures 5.4 and 5.5.



(A) TP (B) TN

(C) FP (D) FN

FIGURE B.1: **Seen data**: probability density functions of the confidence values of all predictions for the *seen data* estimated with kernel density estimation.



(A) TP (B) TN

(C) FP (D) FN

FIGURE B.2: **Unseen data**: probability density functions of the confidence values of all predictions for the *unseen data* estimated with kernel density estimation.

Appendix C

Source code

The source code of this thesis project can be found on GitHub:
https://github.com/delftcrowd/smart_rejector_for_hate_speech

Bibliography

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data* (pp. 77–128). Springer.

Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. *European conference on information retrieval*, 141–153.

Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, *9*, 106363–106374.

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, *40*(7), 64–65.

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 45–54.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *Proceedings of the 26th international conference on World Wide Web companion*, 759–760.

Balayn, A., Yang, J., Szlavik, Z., & Bozzon, A. (2021). Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *ACM Transactions on Social Computing (TSC)*, *4*(3), 1–56.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*(1), 32–68.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. *Proceedings of the 13th international workshop on semantic evaluation*, 54–63.

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of extension*, *50*(2), 1–5.

Casati, F., Noël, P.-A., & Yang, J. (2021). On the value of ml models. *arXiv preprint arXiv:2112.06775*.

Coenen, L., Abdullah, A. K., & Guns, T. (2020). Probability of default estimation, with a reject option. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 439–448.

The constitution [Visited on 11/04/2022]. (n.d.). *The White House*. https://www.whitehouse.gov/about-the-white-house/our-government/the-constitution/

Council of Europe. (n.d.). Hate speech and violence [Visited on 19/01/2022]. *European Commission against Racism and Intolerance (ECRI)*. https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence

Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach. *Science and engineering ethics*, *12*(4), 701–715.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515.

De Stefano, C., Sansone, C., & Vento, M. (2000). To reject or not to reject: That is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *30*(1), 84–94.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edwards, C. (2021). The best of nlp. *Communications of the ACM*, *64*(4), 9–11.

European Commission. (2016). The eu code of conduct on countering illegal hate speech online [Visited on 07/03/2022]. *European Commission*. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Fitzner, K. (2007). Reliability and validity a quick review. *The Diabetes Educator*, *33*(5), 775–780.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, (2020-1).

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4885–4894.

Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 2151–2159). PMLR. https://proceedings.mlr.press/v97/geifman19a.html

Giansiracusa, N. (2021). Facebook uses deceptive math to hide its hate speech problem [Visited on 07/03/2022]. *Wired*. https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/

Gold, M. W. T. H. D., & Zesch, T. (2018). Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments.

Grandvalet, Y., Rakotomamonjy, A., Keshet, J., & Canu, S. (2008). Support vector machines with a reject option. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 468–469.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love" evading hate speech detection. *Proceedings of the 11th ACM workshop on artificial intelligence and security*, 2–12.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International conference on machine learning*, 1321–1330.

Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., & Davis, J. (2021). Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Ingram, M. (2018). Facebook now linked to violence in the philippines, libya, germany, myanmar, and india ["Visited on 07/03/2022"]. *Columbia Journalism Review*. https://www.cjr.org/the_media_today/facebook-linked-to-violence.php

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. (2017). Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, *30*(3), 411–433.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., & Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Lodge, M., Tanenhaus, J., Cross, D., Tursky, B., Foley, M. A., & Foley, H. (1976). The calibration and cross-modal validation of ratio scales of political opinion in survey research. *Social Science Research*, *5*(4), 325–347.

Lodge, M., & Tursky, B. (1979). Comparisons between category and magnitude scaling of political opinion employing src/cps items. *American Political Science Review*, *73*(1), 50–66.

Maddalena, E., Mizzaro, S., Scholer, F., & Turpin, A. (2017). On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, *35*(3), 1–32.

Mashal, M., Raj, S., & Kumar, H. (2022). As officials look away, hate speech in india nears dangerous levels [Visited on 07/03/2022]. *The New York Times*. https://www.nytimes.com/2022/02/08/world/asia/india-hate-speech-muslims.html

McGee, M. (2004). Master usability scaling: Magnitude estimation and master scaling applied to usability measurement. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 335–342.

Moskowitz, H. R. (1977). Magnitude estimation: Notes on what, how, when, and why to use it. *Journal of Food Quality*, *1*(3), 195–227.

Mozur, P. (2018). A genocide incited on facebook, with posts from myanmars military [Visited on 07/03/2022]. *The New York Times*. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131–2167.

Murray, J. (2013). Likert data: What to use, parametric or non-parametric? *International Journal of Business and Social Science*, *4*(11).

Nadeem, M. S. A., Zucker, J.-D., & Hanczar, B. (2009). Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In S. Deroski, P. Guerts, & J. Rousu (Eds.), *Proceedings of the third international workshop on machine learning in systems biology* (pp. 65–81). PMLR. https://proceedings.mlr.press/v8/nadeem10a.html

Norman, G. (2010). Likert scales, levels of measurement and the laws of statistics. *Advances in health sciences education*, *15*(5), 625–632.

Olson, J. S., & Kellogg, W. A. (2014). *Ways of knowing in hci* (Vol. 2). Springer.

Olteanu, A., Talamadupula, K., & Varshney, K. R. (2017). The limits of abstract evaluation metrics: The case of hate speech detection. *Proceedings of the 2017 ACM on Web Science Conference*, 405–406.

Posner, R. A. (1986). Free speech in an economic perspective. *Suffolk University Law Review*, *20*, 1.

Rodriguez, A., Argueta, C., & Chen, Y.-L. (2019). Automatic detection of hate speech on facebook using sentiment and emotion analysis. *2019 international conference on artificial intelligence in information and communication (ICAIIC)*, 169–174.

Roitero, K., Maddalena, E., Demartini, G., & Mizzaro, S. (2018). On fine-grained relevance scales. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 675–684.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2020). Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sayin, B., Yang, J., Passerini, A., & Casati, F. (2021). The science of rejection: A research area for human computation. *arXiv preprint arXiv:2111.06736*.

Schmidt, A., & Wiegand, M. (2019). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*, 1–10.

Social media firms faces huge hate speech fines in germany [Visited on 11/04/2022]. (2017). *BBC News*. https://www.bbc.com/news/technology-39506114

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, *69*(1), 1–25.

Sunstein, C. R. (2019). Does the clear and present danger test survive cost-benefit analysis? *Cornell Law Review*, *104*, 1775.

Tworek, H., & Leerssen, P. (2019). An analysis of germany's netzdg law. *Translantic Working Group*.

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, *1*(3), 283–296.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychologya discussion and suggested template. *Journal of experimental social psychology*, *67*, 2–12.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of the NAACL student research workshop*, 88–93.

Woo, W. L. (2020). Future trends in i&m: Human-machine co-creation in the rise of ai. *IEEE Instrumentation & Measurement Magazine*, *23*(2), 71–73.

Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. *Proceedings of the 21st ACM international conference on Information and knowledge management*, 1980–1984.

Zhu, H., Yu, B., Halfaker, A., & Terveen, L. (2018). Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction*, *2*(CSCW), 1–23.