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. J. Yang
Delft University of Technology

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

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

August 26, 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

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Contents

List of Abbreviations

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
PDF	**p**roability **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 platforms 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.). There are many news articles about events where hate spread on online platforms lead 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 on 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 their progress every year (European Commission, 2016). The most recent report from 2021 stated that Twitter only removed 49.5% of all hateful content on their platform. Facebook is most successful in removing hate speech as they claim 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 hard problem that even large institutions have not solved yet.

Currently, people rely on reactive and proactive content moderation methods to detect hate speech (Klonick, 2018). Reactive moderation is when social media users are flagging (also known as reporting) hateful content (Klonick, 2018). Proactive moderation is either 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 can range from traditional ML methods such as Support Vector Machine or Decision Tree to Deep Learning 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 is simply infeasible due to 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 hateful content is exposed to the users for some time. Proactive automatic moderation using automated detection algorithms allow for large amounts of data to be checked quickly without the involvement of humans. However, these algorithms have shown to be unreliable as they often perform poor on deployment data (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 dataset (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). The authors found that the performance drops significantly when the detection algorithms are trained on one dataset and evaluated on another (Arango et al., 2019).

This thesis research will tackle the problems of proactive moderation by focusing on the concept of *human-machine co-creation* (Woo, 2020), where the advantages of humans (cognitive abilities and ability to make judgements) and machines (automation and performance) are combined. 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 (Woo, 2020). We apply a reject option to some state-of-the-art automatic hate speech detection models. The goal of the reject option is to reject an ML prediction when the risk of making an incorrect prediction is too high and to defer the prediction task to a human (Hendrickx et al., 2021). There are several advantages. First, the utility of the ML model increases as only the most confident (and possibly the most correct) predictions are accepted. Second, less human effort is necessary as the machine handles all prediction tasks, and only a fraction needs to be checked by a human. To the best of our knowledge, ML with rejection has not been used in hate speech detection before.

In this work, we focus on *value-senstive* rejection. There are benefits of accepting correct predictions (positive value) and costs of accepting incorrect or rejecting predictions (negative value). More specifically, we should weigh cost values for false negative (FN), labelling something as non-hateful when it is, and false positive (FP) predictions, labelling something as hateful when it is not, according to the task of hate speech detection and incorporate them in the design of the hybrid human-AI system (Sayin et al., 2021). We will mainly focus on the human values from the perspective of the social media users since they are the most affected by the consequences of hate speech.

The idea of most ML models with rejection is that we reject predictions when the model's confidence is too low. Therefore, we need a metric that measures the total value of ML models with a reject option. We can use the resulting metric to determine when to reject/accept predictions by maximizing the total value. Second, we need to find out how we can define the user-centred values in the context of hate speech detection. We will attempt to retrieve the value ratios since it is hard to determine the absolute cost values in the hate speech domain. By value ratios, we mean, for example, the ratio between an FP and an FN prediction. Therefore, our research questions are as follows:

This leads to the following research questions:

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?

Here comes a list of contributions

Here comes a short description of the structure of the thesis report

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 gender, 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) found 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 that 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. In this project, we focus on hate speech detection as a binary text classification problem where 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

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

with various configurations of word embeddings and two DL models: a convolutional 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 for identifying 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 Support Vector Machine (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 support vector machines (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 (Geifman & El-Yaniv, 2017). 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 their 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 applying the reject option to any existing 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 underlying 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 a value-sensitive 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 both 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 (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) (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 the social media users only.

2.5.1 Quantitative assessment

In this section, we explain the difficulties of defining the values of TP, TN, FP, FN, and rejected predictions in hate speech detection using quantitative measurements. 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 moderate hate speech (Klonick, 2018). So it is infeasible to assess the values of the 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 from the companies' perspective as the stakeholder.

Second, we look at the social media users as the 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.). The authors explained in Posner (1986) 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 difficult (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 commonly used Likert scales as our measurement scale. However, as we will explain in section 2.5.2, Likert scales are unsuitable for retrieving ratio values. Therefore, in section 2.5.2, we will 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. However, there is a lot of 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 that consists of multiple Likert items as interval data and, 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 true zero point (Allen & Seaman, 2007)). However, we need to have ratio data in this project since we want to know the exact value ratios between the TP, TN, FP, FN, and rejected predictions.

Magnitude Estimation

In section 2.5.2, we concluded 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, or duration) and proved 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 judging 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 now 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. 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 in Bard et al. (1996), Maddalena et al. (2017), and McGee (2004), 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 magnitudes estimates are in the range [-100, 100] 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 science 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.

Chapter 3

Value-sensitive rejector

As concluded in the Related work, there is a need for *value-sensitive* metrics for measuring the performance of ML models, 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 rejecting ML predictions in a value-sensitive manner. We do this by taking the values of TP, TN, FP, FN, and rejected predictions into account. Section 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.

3.1 Value-sensitive metric

The idea of rejecting ML predictions using a 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. 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 can derive the subsets of TP, TN, FP, and FN predictions from the complete set of predictions and their 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 gains and, therefore, positive values and V_{fp}, V_{fn}, and V_r are costs and, therefore, negative values. 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 values V_t 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 when 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 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:

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

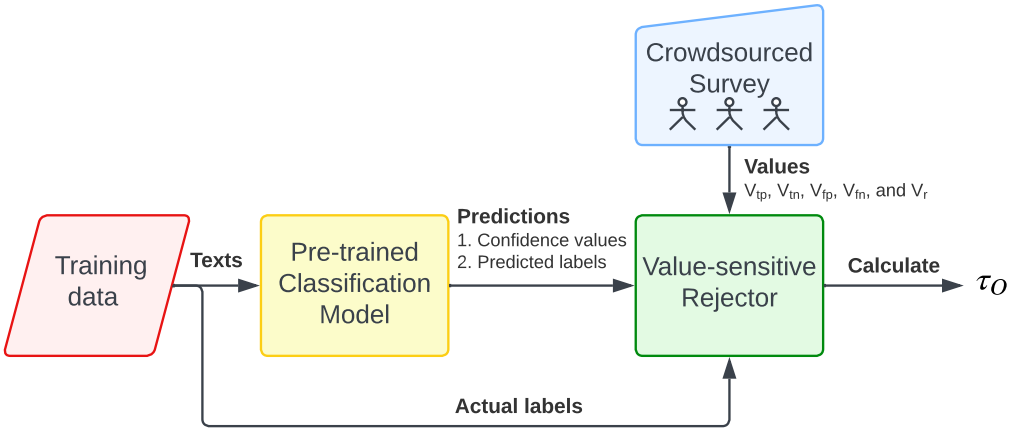

FIGURE 3.1: **Training phase:** flow diagram that visualizes how the value-sensitive rejector calculates the optimal rejection threshold τ_O.

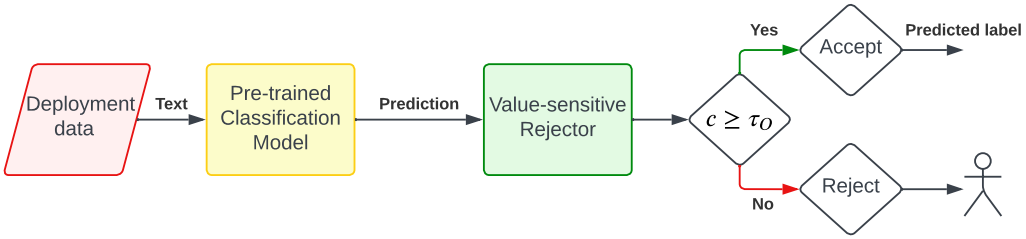

FIGURE 3.2: **Deployment phase:** flow diagram that visualizes how the value-sensitive rejector uses the optimal rejection threshold τ_O and the prediction confidence c to determine when to accept or reject a prediction from unseen data in deployment.

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 Logistich Regression (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 Convolutional Neural Network (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 transformer 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 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.

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

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

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.

3.3.3 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. 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* dataset and the *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 on the seen dataset, similar to the findings of related studies: 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 lower and that the optimal rejection threshold will be higher (meaning that we need to reject more predictions).

3.3.4 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[6] for estimating these PDFs. With KDE, we estimate the PDF by weighing each measured confidence value from a set of predictions using a kernel function, a gaussian density function since it is the

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

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.

3.3.5 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.

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 value ratios from the perspective of the social media user. In section 2.5.2, we found that Magnitude Estimation (ME) seems like a promising technique for estimating these 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. 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, conversely for the FN scenario. 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 the degree of agreement/disagreement using a scale, and we aggregate the answers per scenario 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 and 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 setup of the survey. Finally, in section 4.3, we elaborate on the analysis of the survey results.

4.1 Hypothesis

Before we conducted the survey experiment, we listed several hypotheses about the value ratios and the ME method:

- **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 has a more negative

impact on 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 and, therefore, we think that 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.
- **We hypothesize that Magnitude Estimation (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

In this section, we discuss 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 provides more response freedom than course-grained Likert scales and 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 Likert scales (Roitero et al., 2018). Therefore, we will 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 magnitudes estimates are in the range $[-100, 100]$ while maintaining the ratio properties.

4.2.3 Design

We will list all independent, dependent, confounding, and control variables analyzed in our experiment in this section.

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. We decided to explain that the hateful posts are ranked lower since we found significant disagreement among participants after the pilot survey in which we explained that SocialNet removes hateful posts. 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 how we should deal with it.
- **Age** People of different ages might have different perceptions and definitions of hate speech and how we should deal with it.
- **Educational level** People with different educational levels might have different perceptions and definitions of hate speech and how we should deal with it.
- **Gender** According to Gold and Zesch (2018), there is no significant difference in how men and women perceive hate. However, we still report gender 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 Magnitude Estimation 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 reliability, validity and the value ratio of TP, TN, FP, FN, and rejection scenarios.

- **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 estimates 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, recruit the participants, and 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 (CI) and 10% Margin of Error (MoE) for this study. So for 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 CI and MoE values. We choose 10% MoE since we have a limiting 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 decide whether reducing the MoE by increasing the number of participants is possible. For the pilot survey, we use 24 participants. Therefore, in total we will 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 will use the Prolific platform for recruiting online participants for the survey study. We will use the following inclusion criteria for our participants:

- 18 years of age and older since we show offensive language in the experiment.
- 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 will be 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 will include two Instructional Manipulation Checks to check if the user pays attention to the survey[3].

[1] https://datareportal.com/reports/digital-2021-october-global-statshot
[2] https://prolific.co/pricing
[3] https://researcher-help.prolific.co/hc/en-gb/articles/360009223553

- 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.
- Participants who do fail the ME training phase. For example, by providing random values that do not make any sense.

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 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, 2012). The term frequencies are multiplied with 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 to set 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

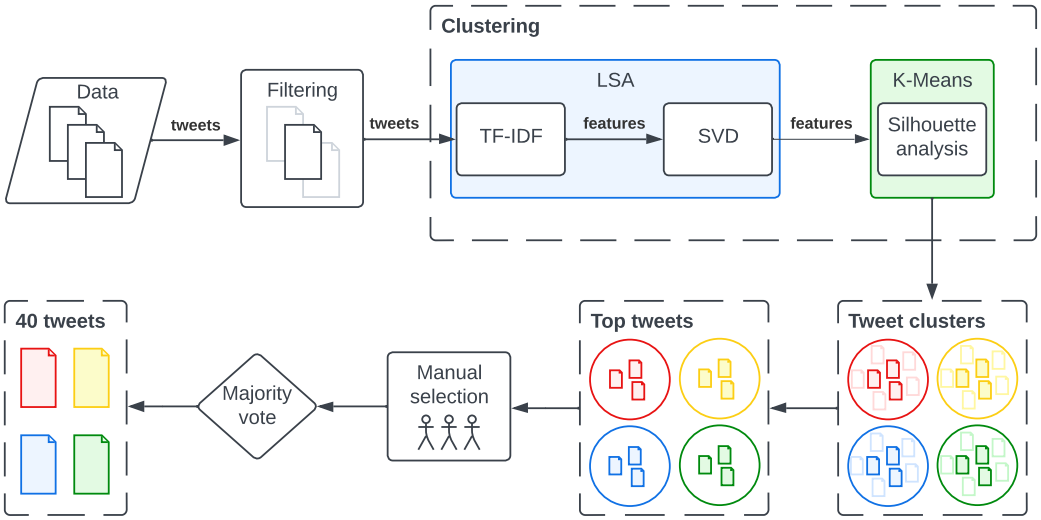

FIGURE 4.1: Flow diagram that visualizes how we perform content analysis to cluster and select the tweets for our survey study.

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.

4.2.6 Procedure

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 will expect from them and to explain the structure of the survey. Using the ME scale, we first present a warm-up task where the participants need to estimate different line lengths to get familiar with using the ME scale. Then, we randomly present 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 will inform the participants not to put identifiers in their answers. Refer to Appendix A for all presentation texts, the informed consent, and some scenario examples.

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

Step 1: provide informed consent

- Show the informed consent (with checkboxes for giving consent).
- Proceed to the next step only for the participants who give consent.

Step 2: introduction

- Show introductory text about what is expected from the participant.
- We split all participants up into two groups.
- The first group first uses the ME scale to rate all scenarios.
- The second group first uses the 100-level scale to rate all scenarios.
- Explanation of the scale.

Step 3: two attention checks

- Two simple attention checks where we ask the participant to select one option (shuffled through all scenarios).

Step 4a: ME practice phase (when ME is used)

- To let participants learn how to use ME, we first run a practice phase where we shuffle and present 5 different line lengths.
- Each participant needs to estimate the line length using any positive value.

Step 4b: all scenarios using the scale (ME or 100-level)

- Show 40 different scenarios in random order: 8 TP, 8 TN, 8 FP, 8 FN, and 8 rejection.

Step 5: finish

- Show a thank you message and redirect the users to Prolific to complete the task.

4.3 Analysis

First, we calculate the value ratios of 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 value ratios of 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 will 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:

$$V_{tp}^{ME} = \frac{1}{n}\sum_{i=1}^{n} r_{i,tp}^{ME}$$ where n is the total number of all participants for TP scenarios and $r_{i,tp}^{ME}$ is the median response value of TP question number i rated with the ME scale.

$$V_{tp}^{100} = \frac{1}{n}\sum_{i=1}^{n} r_{i,tp}^{100}$$ where n is the total number of all participants for TP scenarios and $r_{i,tp}^{100}$ is the median response value of TP question number i rated 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 will 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 will 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 the 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 will 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 guarantee 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 are still curious if there are any significant statistical differences between groups of participants with different demographic characteristics as 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 four variables: nationality, age, student (whether they are still a student or not), and gender. We have multiple groups (more than two) for the variables nationality and age (different age intervals) and two groups for the variables student and gender. We apply either analysis of variance (ANOVA) (parametric) or Kruskal-Wallis (non-parametric) when we have more than two groups and apply an unpaired two-samples 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 (normal distribution of the error), 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 of our survey study Independently. Although ANOVA and the t-test can be robust to violations of the homogeneity of variances and the normality assumptions (Howell, 2012), we decide to use the non-parametric alternatives instead to prevent us from getting unreliable results.

Chapter 5

Results

5.1 Survey study

5.1.1 Value ratios

Explain that the metric conditions still hold if we set V_{tp} and V_{tn} to 0

5.1.2 Reliability

5.1.3 Validity

5.1.4 Demographics

5.2 Value-sensitive rejection

Chapter 6

Discussion

Answer research questions

Discuss future work

6.1 Survey study

6.2 Value-sensitive rejection

6.3 Implications

Explain that Olteanu et al., 2017 claims that we need more human-centred metrics instead of abstract metrics such as precision and we agree with that by introducing our own human-centred metric

6.4 Limitations

Hate speech is difficult domain as there tend to be a lot of disagreement between people about what is considered hate speech and what not. Ross et al. (2017) found low Krippendorff alpha values in a hate speech survey. So our findings are in line with theirs.

Explain limitations of the metric and the survey study

The rejection threshold is calculated using the test set. This test set needs to be as realistic as possible. Furthermore we need to have calibrated models since we rely purely on the confidence values. This is also hard to realize. Temperature scaling can help, but it is still limited.

6.5 Recommendations

Magnitude Estimation seems promising for future research in HCI.

Personal and demographic characterisitcs might have a big impact. So further analysis on those aspects seem relevant.

Perhaps we can train ML models using the values of TP, TN, FP, FN, rejection in an integrated rejector. So we train the ML model and the rejector simultanously using the values from the survey. So then during training, the FN predictions are punished more than FP predictions.

Chapter 7

Conclusion

Appendix A

Survey

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

A.1 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.2 Introduction

A.2.1 Short introduction ME

- 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.2.2 Short introduction 100

- 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.2.3 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.3 Scales

A.3.1 100-level scale explanation

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.3.2 ME scale explanation (inspired by 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 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.5 Examples

A.5.1 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.5.2 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.5.3 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.

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

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*.

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

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.

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.