

DELFT UNIVERSITY OF TECHNOLOGY

MASTERS THESIS

Value-Sensitive Rejection of Machine Decisions 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

July 11, 2022

DELFT UNIVERSITY OF TECHNOLOGY

Abstract

Electrical Engineering, Mathematics and Computer Science
Software Technology

Master of Science

Value-Sensitive Rejection of Machine Decisions 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

Abstract	iii
Acknowledgements	v
1 Introduction	1
2 Related work	3
2.1 Hate speech detection	3
2.2 Machine Learning models with rejection	4
2.3 Value-sensitive hate speech detection	4
2.3.1 Objective values	5
2.3.2 Subjective costs	6
3 Value-sensitive rejection	7
3.1 Metric	7
3.2 Hate speech detection models	7
3.2.1 Logistic Regression	7
3.2.2 CNN	7
3.2.3 DistilBERT	7
3.2.4 Calibration	7
4 Survey study	9
4.1 Research question and hypothesis	9
4.2 Method	9
4.2.1 Scales	9
4.2.2 Normalization	9
4.2.3 Design	9
Independent variables	9
Confounding variables	9
Control variables	9
Dependent variables	9
4.2.4 Planned sample	9
Participant Inclusion and Exclusion Criteria	9
Participant Compensation	9
Sample size	9
4.2.5 Materials	9
Survey tool	9
Data	9
4.3 Analysis	9
4.3.1 Validation	9
4.3.2 Reliability	9

5	Results	11
5.1	Survey	11
5.1.1	Value ratios	11
5.1.2	Reliability	11
5.1.3	Validity	11
5.2	Metric	11
6	Discussion	13
7	Conclusion	15
	Bibliography	17

List of Abbreviations

Add abbreviations

List of Symbols

Add symbols

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 both humans (cognitive abilities and ability to make judgements) and machines (automation and performance) are combined. So humans and machines should work together to detect hate speech. ML models should detect hateful content automatically and humans should make the final decisions (*human-in-the-loop*) when the model is not confident enough (Woo, 2020). Here come ML models with a reject option in place. 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 increases as only the most confident (and possibly the most correct) predictions are accepted. Second, less human effort is necessary as the machine is handling 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-sensitive* rejection. There are gains of accepting correct predictions (positive value) and costs of accepting incorrect or rejecting predictions (negative value). We should weigh these values according to the task of hate speech detection and incorporate them in the design of the human-AI collaboration framework Sayin et al., 2021. We will mainly focus on the user-centred value since the social media users are the most affected by the consequences of hate speech.

Continue here

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 utility of ML models with a reject option. We can use the resulting metric to determine when to reject/accept predictions by maximizing the utility 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 come up with the absolute cost values in the hate speech domain. By value ratios, we mean to figure out, for example, the ratio between an FP and an FN prediction. Therefore, our first sub-research question is as follows:

This leads to the following research questions:

RQ How can we maximize the value of Machine Learning models in hate speech detection using a reject option?

- **SRQ1** How can we measure the utility of Machine Learning models with a reject option?
- **SRQ1** 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

This section gives some background information about ML with rejection, hate speech detection, and unknown unknown detection

2.1 Hate speech detection

In this section, we first briefly introduce the definition of hate speech and the challenges of detecting it. Then, we provide an overview of some methods commonly used for automatically detecting hate speech.

Different types of online conflictual languages exist, such as cyberbullying, offensive language, toxic language, or hate speech, and come with different 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’s mostly focused on specific target groups or individuals (Balayn et al., 2021). But there exist many more definitions of hate speech in literature mainly because there is a lot of discussion about what is considered hate speech and what not. From several studies, it appears that there is low agreement among humans when it comes to 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 20 tweets. They also found that the inter-rater reliability value does not increase when showing a definition of hate speech to the human annotators before the survey. Waseem (2016) found a slight increase in the inter-rater reliability when considering annotations of human experts only, but it remained low overall. This low agreement makes sense since there are many differences in people’s personalities and backgrounds. Therefore, hate speech detection is challenging, especially in computer science, since we have to be careful with bias. Most annotated hate speech datasets that are publicly available contain bias. 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 algorithms are likely to become biased in their predictions if they are trained on biased datasets. Bias becomes most notable when applying pre-trained classification algorithms to new and unseen data in deployment. For example, Gröndahl et al. (2018) and Arango et al. (2019) report significant drops in F1 scores when training a hate speech detection model on one dataset and evaluating it on another. These results

further strengthen our stance that hate speech detection cannot and should not be done by machines only but rather by a human-in-the-loop approach.

2.2 Machine Learning models with rejection

Explain the different architectures of ML with rejection

Explain the different types of confidence metrics

Explain the original metric from De Stefano

Provide examples of ML models with rejection from other domains

2.3 Value-sensitive hate speech detection

Explain how we could do value analysis in hate speech detection and why some methods do not work (such as expressing the values in money or time) and which methods might work (such as using surveys for retrieving subjective values). Also, explain what Magnitude Estimation is and provide examples of studies that used it to retrieve subjective judgements.

In this research, we focus on Machine Learning models with a reject option. The decision to accept or reject predictions depends heavily on the context. We argued in the **Introduction** that this decision should depend on the costs of incorrect predictions and the gains of correct predictions. We can express the costs of incorrect (FP and FN predictions) and rejected predictions as negative values. The gains of correct predictions (TP and TN predictions) as positive values. In some domains, we can define these values in money or time. For example, suppose there is a factory that uses a camera and an ML model to check if a package is damaged or not. Using an ML model will save the company time since these packages no longer have to be inspected manually by humans. However, the ML model could be incorrect sometimes. For example, a customer of the factory received a damaged package, while the ML model did not detect any damage. Fixing this issue could cost the factory money. At the same time, the factory could prevent these cases by rejecting the low confidence predictions from the ML model. For example, the ML model predicted with low confidence that a package did not contain any damage. An employee can then inspect it to prevent the customer from receiving a damaged one. Manually checking the rejected ones costs the factory a fraction of the time/money compared to the first situation. In this example, we can easily express the values of FP, FN, TP, TN, and rejections in time/money spent/saved.

However, it is not evident to express these values in the hate speech domain. Two stakeholders can be considered in the design of a smart rejector: the social media company and its users. We mainly focus on the users in this research since they will be affected the most by hate speech.

In this section, we will look at the related work to get an understanding of how we could retrieve the value ratios in hate speech detection. The goal is to retrieve ratios between rejection, FP, FN, TP, and TN cases. We would like to know whether an FN is, for example, two times worse compared to an FP. The main challenge is to express all values using a single unit. We could take two directions. First, we could define the values using an objective measure, such as time or money spent/saved. Second, we could define the values subjectively, e.g. by analyzing people's stance

towards the consequence of incorrect predictions in hate speech detection. In the next two sections, we discuss the relevant related work in both directions.

2.3.1 Objective values

In this section, we explain the difficulties of defining the values using objective measurements. We do this by looking at some related work. 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 taking the moderator's salary into account. We could also argue that the value of a TP and a TN is equal to the negative value of rejection since we saved human effort by letting the ML model correctly predict whether something is hateful or not. 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 in terms of money or time?

First, we look at the social media company as a stakeholder. As we explained in the previous section, the values of rejection, TP, and TN can be determined. So the values of FP and FN are yet to be defined. However, most social media companies are not transparent in how they moderate hate speech (Klonick, 2018). So we do not have clear insights into the costs for these companies. There do exist country-specific 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 cases is even more difficult. It is unclear how filtering out too much content would affect the company (apart from many annoyed users whose freedom of speech is violated). Therefore, we abstain from estimating the values from the perspective of these companies.

Second, both FP and FN predictions have consequences on the users as the stakeholder. 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 that 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 need to figure out how we should 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 come up with 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, for example, 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). For example, 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 not ever result in actual acts of violence (Sunstein, 2019). Therefore, cost values using objective measurements are often case-specific and cannot be defined generically. There is a nonquantifiable risk that acts of violence will happen in the unknown future (Sunstein, 2019). But suppose we do know this probability, then there are still too many uncertainties. To calculate the actual costs of hate speech (in our case: to accept 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 in the hate speech context (Sunstein, 2019).

Therefore, we believe that using objective measurements, such as money, is not realistic for generically expressing the cost values in our project for both stakeholders.

2.3.2 Subjective costs

Chapter 3

Value-sensitive rejection

3.1 Metric

Explain and proof our modified version of the metric from De Stefano

3.2 Hate speech detection models

3.2.1 Logistic Regression

3.2.2 CNN

3.2.3 DistilBERT

3.2.4 Calibration

Explain what model calibration is and why it's necessary

Chapter 4

Survey study

4.1 Research question and hypothesis

4.2 Method

4.2.1 Scales

4.2.2 Normalization

4.2.3 Design

Independent variables

Confounding variables

Control variables

Dependent variables

4.2.4 Planned sample

Participant Inclusion and Exclusion Criteria

Participant Compensation

Sample size

4.2.5 Materials

Survey tool

Data

4.3 Analysis

4.3.1 Validation

4.3.2 Reliability

Chapter 5

Results

5.1 Survey

5.1.1 Value ratios

5.1.2 Reliability

5.1.3 Validity

5.2 Metric

Chapter 6

Discussion

Answer research questions

The rejection threshold is calculated using the test set. This test set needs to be as realistic as possible.

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. Most datasets are binary labeled but perhaps it's better that hate speech datasets use an ordinal scale to define how hateful a text sample is.

Explain difficulties in coming up with numerical cost/gain values of (in)correct predictions and rejections

Discuss future work

Chapter 7

Conclusion

Bibliography

- 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.
- 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.
- 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.
- 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., Warmley, 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.
- 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/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4), 1–30.
- 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/>
- 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.
- 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*.
- 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.

- 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>
- 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.
- 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.
- 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*.
- Sayin, B., Yang, J., Passerini, A., & Casati, F. (2021). The science of rejection: A research area for human computation. *arXiv preprint arXiv:2111.06736*.
- 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>
- Sunstein, C. R. (2019). Does the clear and present danger test survive cost-benefit analysis? *Cornell Law Review*, 104, 1775.
- 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.