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 17, 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

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 model 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-senstive* 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.

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 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 reject predictions of Machine Learning models in hate speech detection in a value-sensitive manner?

- **SRQ1** How can we measure the value 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 briefly defines hate speech and why it is such a challenging topic to tackle, especially in computer science. Then, we give an overview of some algorithms we can use to detect hate speech automatically. We provide examples of rejecting classification model predictions and discuss the main challenges of assessing the values of (in)correct and rejected predictions in the context of hate speech detection. Finally, we discuss why standard metric, such as accuracy, are unsuitable to evaluate hate speech detection models and why we need human-centred metrics instead.

2.1 Challenges of hate speech detection

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). However, many more definitions exist in literature mainly because people differ on what is considered hate speech and what is not. Balayn et al. (2021) identified the mismatch between the formalisation 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 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 beforehand. Waseem (2016) found a slight increase in the inter-rater reliability when considering annotations of human experts only, but it remained low overall. 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. Annotating hate speech datasets is challenging because social media data follows a skewed distribution since there are many more neutral social media posts than hateful ones (Fortuna & Nunes, 2018). 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 will likely become biased in their predictions if we train them 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. 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 detection 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 Detection algorithms

There is an increasing academic interest in the automatic detection of hate speech since the topic has become more relevant, as explained in the Introduction. We will list the state-of-the-art Natural Language Processing (NLP) techniques for automatic hate speech detection from literature. First, we will list the different features used in the classification approaches. Then, we will list the currently used classification algorithms ranging from supervised to unsupervised learning.

Several excellent surveys outlined the classification approaches from literature (Fortuna & Nunes, 2018; Schmidt & Wiegand, 2019). Commonly used features are bag-of-words (BOW) (Greevy & Smeaton, 2004), character/word N-grams (Waseem & Hovy, 2016), lexicon features (e.g. using a word blacklist containing offensive slurs) (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 the topic distributions from an LDA analysis and a lexicon feature. 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.

Most hate speech-related studies use supervised learning techniques that range from traditional ML to deep learning 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 traditional ML techniques such as Naive Bayes, Decision Trees, and Random Forests. Badjatiya et al. (2017)

[1] https://fasttext.cc/

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

experimented with various configurations of word embeddings and two deep learning models: a convolutional neural network (CNN) and a long short-term memory (LSTM) model. They found that CNN performs better than LSTM and that using pre-trained word embeddings such as GloVe does not result in better classification performance than using random embeddings. Rodriguez et al. (2019) use the unsupervised learning method, K-means clustering, to cluster social media posts to identify topics that potentially promote hate speech.

Provide example of BERT model for hate speech detection

2.3 Machine Learning models with rejection

Several studies promoted the concept of rejecting ML predictions when the risk of producing an incorrect prediction is too high (Hendrickx et al., 2021; Sayin et al., 2021; Woo, 2020). Hendrickx et al. (2021) identified three architectures for rejecting ML predictions: the *separated*, the *integrated*, and the *dependent* rejector. The separated rejector shares no information with the classification model and decides beforehand whether a data sample needs to be handled by the classification model or by a human (Hendrickx et al., 2021). The integrated rejector is one model that is a combination of a classification model and a rejector (Hendrickx et al., 2021). The 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 in the context of ML using one of the architectures mentioned above (Coenen et al., 2020; Geifman & El-Yaniv, 2017, 2019; Grandvalet et al., 2008; Nadeem et al., 2009).

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 that were 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 a deep learning (DL) model by including a selection function in the last layer of the DL model.

In this work, we apply the dependent architecture since it supports any existing classification model (Hendrickx et al., 2021). The most relevant work is from De Stefano et al. (2000) since their confidence metric takes the value of (in)correct and rejected predictions) into account. While they experimented with a range of different cost values, we go further by employing a value-sensitive approach, which determines cost values based on how users feel regarding machine decisions using a survey study with crowd workers. Thus, we obtain a threshold that captures the implications of machine decisions from a human perspective.

2.4 Value assessment

Explain what we mean by value and why we should integrate it into the design of a hate speech detection system.

The authors of Fjeld et al. (2020) outlined 8 principles of AI systems including fairness and discrimination (e.g. algorithmic bias), human control of technology (e.g. AI system should request help from the human user in difficult situations), and promotion of human values (we should integrate human value in the AI system).

We need to weigh the value of (in)correct and rejected predictions into account in the design of a hybrid human-AI system (Sayin et al., 2021)

Value sensitive Design (VSD) from Umbrello and Van de Poel (2021) state that we can translate values such as freedom of bias into the design of a system. So example is a tax system that needs to detect fraud. If etniticity bias can be introduced by using postal codes, than we can exclude the postal code variable from the learning algorithm. We for example have the value reliability. The AI is not reliable sometimes, so we use reject option in our design to make the use of AI more reliable/. The authors say that we not only want to optimize the tax algotihm in terms of effectiveness (rate of fraud detection) but also in terms of fairness (presenting non-biased selection of cases. So the same holds in our case, we want to optimize in using the AI as much as possible but also want to take mistakes into account (in the end we still need to invovle humans to check specific cases).

The Value-design algorithm paper from Zhu et al. (2018) describes 5 steps of algorithm design. We focus our approach on theirs by inspecting the stakeholders. And then assessing the stakeholder values to take this into account in the algoithm design.

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.

2.4.1 Objective assessment

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 classification 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.

2.4.2 Subjective assessment

- Focus on subjective values of users - Not companies

2.5 Evaluation metrics

Most hate speech-related studies evaluate their classification methods using standard *machine* metrics such as accuracy, precision, recall, or F1. Evaluation of classification models with a reject option is often done by analyzing the accuracy and the coverage of the classification model. Nadeem et al. (2009) proposed the use of accuracy-rejection curves to plot the trade-off between accuracy and coverage so that different classification models with a reject option can be compared. Röttger et al. (2020) and Olteanu et al. (2017) recognized these shortcomings of machine metrics such as accuracy and found a gap in the evaluation of hate speech detection systems. Röttger et al. (2020) found that it is 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 different criteria, such as the ability to cope with spelling variations or detect neutral content containing slurs (Röttger et al., 2020). Our approach is different since we focus on measuring the performance 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 detection algorithms. They found that the perceived value varies for fixed machine performance measurements, such as precision, and that it depends 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). Our work aligns with theirs since we also create a human-centred metric for evaluating hate speech detection systems with a reject option.

"proposing new metrics for evaluating machine learning systems that incorporate value parameters (Casati et al., 2021)"

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 study

5.1.1 Value ratios

5.1.2 Reliability

5.1.3 Validity

5.2 Value-sensitive rejection

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

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

6.5 Recommendations

Chapter 7

Conclusion

Bibliography

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

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.

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.

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.

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

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/

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. https://proceedings.neurips.cc/paper/2008/file/3df1d4b96d8976ff5986393e8767f5 Paper.pdf

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.

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.

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

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.

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*.

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

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

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

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.