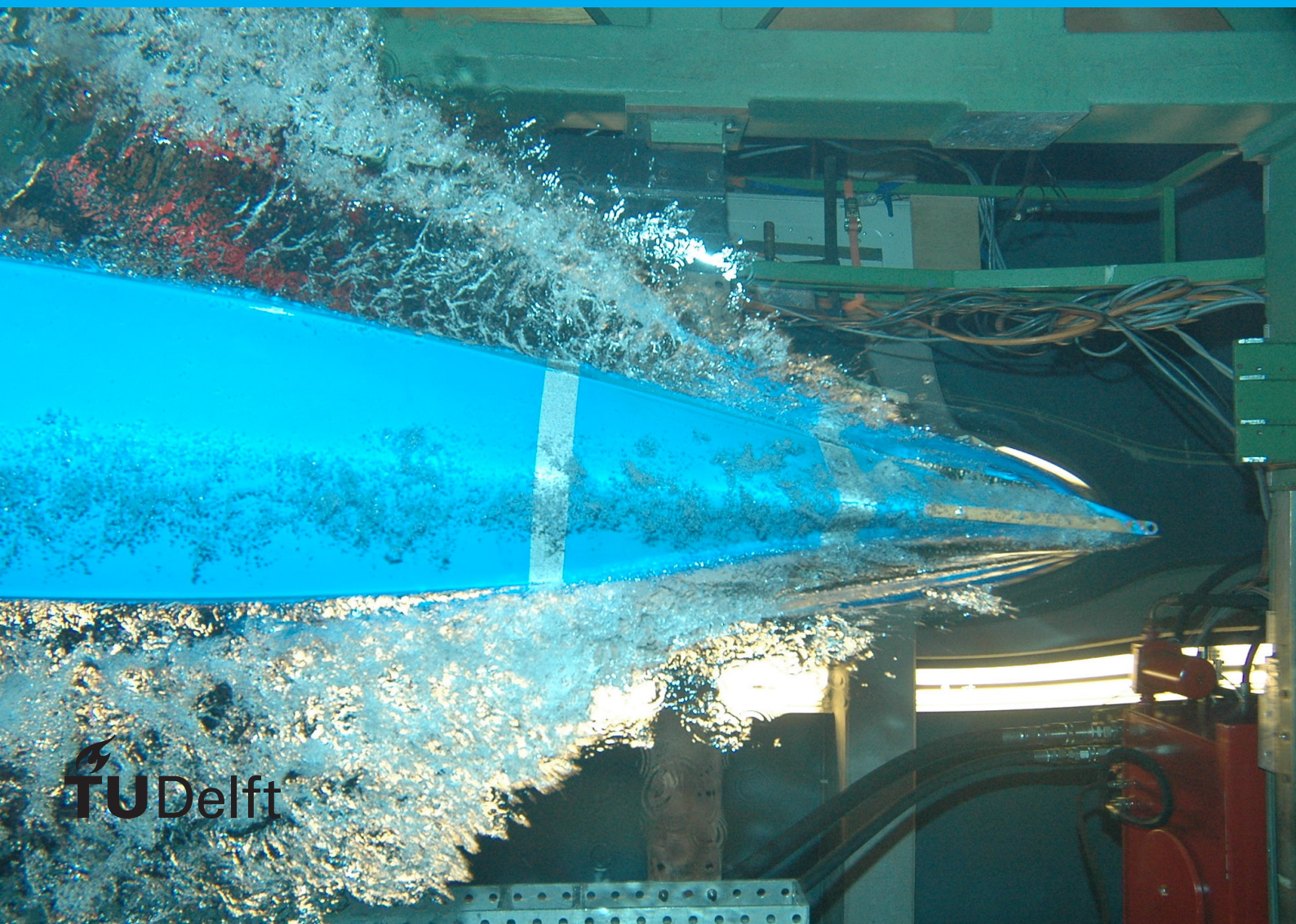
Building a smart rejector for detecting hate speech

Philippe Lammerts



Building a smart rejector for detecting hate speech

by

Philippe Lammerts

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday January 1, 2013 at 10:00 AM.

Student number:	4563182	
Project duration:	September 17, 2021 – TBD	
Thesis committee:	Prof. dr. ir. G.J.P.M. Houben,	TU Delft, thesis advisor
	Dr. J. Yang,	TU Delft, daily supervisor
	Dr. Y-C. Hsu,	TU Delft, co-daily supervisor

TU Delft

Abstract

Acknowledgements

Philippe Lammerts
Delft, January 2013

Contents

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 [1, 6]. There are many news articles about events where hate spread on online platforms lead to acts of violence [13, 16–18]. 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 [18]. 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 [3]. This Code of Conduct requests companies to prohibit hate speech and report their progress every year [3]. 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 [3]. However, one article found in internal communication from Facebook that this percentage is much lower, around 3-5% [10]. 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 [14]. Reactive moderation is when social media users are flagging (also known as reporting) hateful content [14]. Proactive moderation is either done automatically using detection algorithms or manually by a group of human moderators [14]. 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 [6, 9]. These algorithms can range from traditional ML methods such as Support Vector Machine or Decision Tree to Deep Learning algorithms [9].

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 [6]. 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 [6, 11]. 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 [11]. Furthermore, one paper found that most research in hate speech detection overestimates the performance of the automated detection methods [5]. The authors found that the performance drops significantly when the detection algorithms are trained on one dataset and evaluated on another [5].

This thesis research will tackle the problems of proactive moderation by creating a *human-machine co-creation* [26] system that combines the advantages of both humans (cognitive abilities and abil-

ity to make judgements) and machines (automation and performance). A system where humans and machines work together to detect hate speech. This system is a *machine-assisting-human* system where ML models help humans detecting hateful content automatically and where humans can make the final decisions (*human-in-the-loop*) when the model is not confident enough [26]. 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 [12]. 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 predictions 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. Therefore, the goal of this thesis project is to build the first *smart rejector for detecting hate speech*. This leads to our main research question:

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

The idea of most ML models with rejection is that we reject predictions when the model's confidence is too low. However, we need to tackle two underlying problems to be able to answer our main research question. First, we need to figure out when the ML model is not confident enough by measuring the utility of the reject option according to the task of hate speech detection. Second, we need to determine how we can detect the high confidence errors.

The first problem is about the trade-off between how much we trust the predictions produced by the ML model and how much we involve humans to make the judgements. There are gains of accepting correct predictions, costs of accepting incorrect predictions, and costs of rejecting samples. More specifically, we should weigh the 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 [22]. So, we first 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 these 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:

SRQ1 How can we determine when the Machine Learning model is not confident enough?
- **SRQ1.1** How can we measure the utility of Machine Learning models with a reject option?
- **SRQ1.2** How can we determine the value ratios between rejections and True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions?

The second problem is about detecting the low and high confidence errors. These errors are also called *unknown (un)knowns* [15]. When we would only rely on the confidence of the ML model to determine when to reject/accept predictions, then we would accept a subset of incorrect predictions with high confidence, and we would reject a subset of correct predictions with low confidence. We need to find a way to recognize these unknown (un)knowns. Once detected, we can reject the unknown unknowns so that a human moderator makes the final judgement and accept the unknown knowns to save the human moderator extra work. Doing so would further improve the utility of our smart rejector for detecting hate speech. So, our second sub research question is as follows:

I will address the second sub research question only if there is enough time left.

SRQ2 How can we detect the unknown (un)knowns?

Here comes a list of contributions

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

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.

There exist different types of online conflictual languages, such as cyberbullying, offensive language, toxic language, or hate speech, and all come with different definitions from domains such as psychology, political science, or computer science [6]. 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” [6, 8]. It is different from other types of conflictual languages since it’s mostly focused on specific target groups or individuals [6]. But there exist many more definitions of hate speech in literature and the main reason for this is that 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 [9, 21, 24]. Ross et al. [21] 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 [24] found a slight increase in the inter-rater reliability when considering annotations of human experts only but it remained low overall. This low agreement can be explained since there exist many differences in people’s personalities and backgrounds. This makes hate speech detection a challenging task, especially in computer science since we have to be very careful with introducing bias. For example, most annotated hate speech datasets that are publicly available contain bias. Datasets such as [25] or [7] collected their data using specific keywords which can introduce *sample retrieval* bias and annotated their data using only three independent human annotators which might result into *sample annotation* bias [6]. Automated classification algorithms are likely become biased in their predictions as well when they are trained on biased datasets. The effects of bias become mostly visible when we are applying these pre-trained classification algorithms to new and unseen data in deployment. For example, Gröndahl et al. [11] and Arango et al. [5] report significant drops in F1 scores when training a hate speech detection model on one dataset and evaluating it on another. These results further strengh our stance that hate speech detection cannot and should not be solved 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. Cost assessment in hate speech detection

Explain how we could do cost analysis in hate speech detection and why some methods do not work (such as expressing the costs in money or time) and which methods might work (such as using surveys for retrieving subjective costs). 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 consequences of incorrect predictions and the gains of correct predictions. We can express the cost of incorrect predictions as the cost of an FP and an FN. The gains as the gain of a TP and a TN. In some domains, we can define these gain/cost 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 cost/gain values of FP, FN, TP, TN, and rejections in time/money spent/saved.

However, it is not evident to express these costs 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 relative cost values in hate speech detection. The goal is to retrieve relative cost values for 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 the cost values using a single unit. We could take two directions. First, we could define the costs using an objective measure, such as time or money spent/saved. Second, we could define the costs 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 costs

In this section, we explain the difficulties of defining the cost values using objective measurements. We do this by looking at some related work. We can retrieve the cost 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 gain of a TP and a TN is equal to the inverse cost 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 costs 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 costs of rejection, TP, and TN can be determined. So the costs of FP and FN are yet to be defined. However, most social media companies are not transparent in how they moderate hate speech [14]. 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 [4]. However, this is location-specific, and it is unclear how this applies to individual cases of hate speech. Defining the cost 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 costs 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 [19]. Having too many FN predictions might harm individuals or even result in acts of violence [1]. Therefore, we need to figure out how we should weigh the costs 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 gains/costs 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 [20]. The First Amendment restricts the government from creating laws that could, for example, violate Freedom of Speech [2]. The authors explained in [20] 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 [20]. The proposed economic model from the paper, for example, includes the cost of harm and the probability that speech results in violence [20]. However, the authors do not elaborate on how we can define the probability and cost values. Another paper did speculate on this topic by explaining why doing a cost-benefit analysis of free speech is almost impossible [23]. The authors explained that there are too many uncertainties [23]. We can assume that there are costs and benefits of free speech, but it is too difficult to quantify them [23]. For example, terrorist organizations use free speech to recruit people and call for acts of violence online [23]. At the same time, most other hateful posts will not ever result in actual acts of violence [23]. 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 [23]. 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 [23]? The authors claim that analyzing the benefits of free speech is even more difficult [23]. They conclude their work by saying that there are too many problems to empirically evaluate the costs and benefits in the hate speech context [23].

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

2.4. Unknown (un)known detection

Give examples of existing unknown (un)known detection methods from literature

3

Methods

3.1. Hate speech detection

3.1.1. Model

Explain the model's architecture using the original paper from Agrawal and Awekar

3.1.2. Calibration

Explain what model calibration is and why it's necessary

3.2. Cost-utility metric

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

3.3. Cost assessment

Explain how we retrieve the cost values that are used in the cost-utility metric

3.4. Unknown (un)knowns

Explain method used for detecting unknown (un)knowns in the smart rejector

4

Implementation

4.1. System architecture

Explain design of the smart rejector and how the different methods are combined

4.2. Phases

4.2.1. Training

The system will probably support a training and deployment phase. Explain here the training phase of the smart rejector. During this phase, the preparations are done for training the model, determining the optimal rejection threshold, and preparing things for detecting the unknown unknowns

4.2.2. Deployment

Explain how the smart rejector works in the wild and is detecting hate speech in new unlabelled data.

5 Evaluation

5.1. Survey

5.1.1. Setup

Describe the experimental setup

5.1.2. Method

Explain the method for retrieving the cost values for hate speech detection

5.1.3. Results

The low inter-rater reliability is not surprising since this is in line with the results from Ross et al. [21], Waseem [24].

5.2. Smart rejector

5.2.1. Setup

Describe the experimental setup

5.2.2. Method

Explain which experiments are conducted

Explain how the results are analyzed. Things to consider: Accuracy-Rejection curves, accuracy of accepted predictions, rejection rates, acception rates

5.2.3. Results

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

7

Conclusion

Bibliography

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

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

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

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

[5] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54, 2019.

[6] Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 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, 2021.

[7] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, 2017.

[9] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

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

[11] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. All you need is ”love” evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12, 2018.

[12] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*, 2021.

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

[14] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131:1598, 2018.

[15] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. Towards hybrid human-ai workflows for unknown unknown detection. In *Proceedings of The Web Conference 2020*, pages 2432–2442, 2020.

[16] Mujib Mashal, Suhasini Raj, and Hari Kumar. As officials look away, hate speech in india nears dangerous levels. *The New York Times*, Feb 2022. URL `https://www.nytimes.com/2022/02/08/world/asia/india-hate-speech-muslims.html`. Visited on 07/03/2022.

[17] Paul Mozur. A genocide incited on facebook, with posts from myanmar's military. *The New York Times*, Oct 2018. URL `https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html`. Visited on 07/03/2022.

[18] Karsten Müller and Carlo Schwarz. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167, 2021.

[19] Alexandra Olteanu, Kartik Talamadupula, and Kush R Varshney. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 405–406, 2017.

[20] Richard A Posner. Free speech in an economic perspective. *Suffolk University Law Review*, 20: 1, 1986.

[21] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[22] Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. The science of rejection: A research area for human computation. *arXiv preprint arXiv:2111.06736*, 2021.

[23] Cass R Sunstein. Does the clear and present danger test survive cost-benefit analysis? *Cornell Law Review*, 104:1775, Nov 2019.

[24] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.

[25] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

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