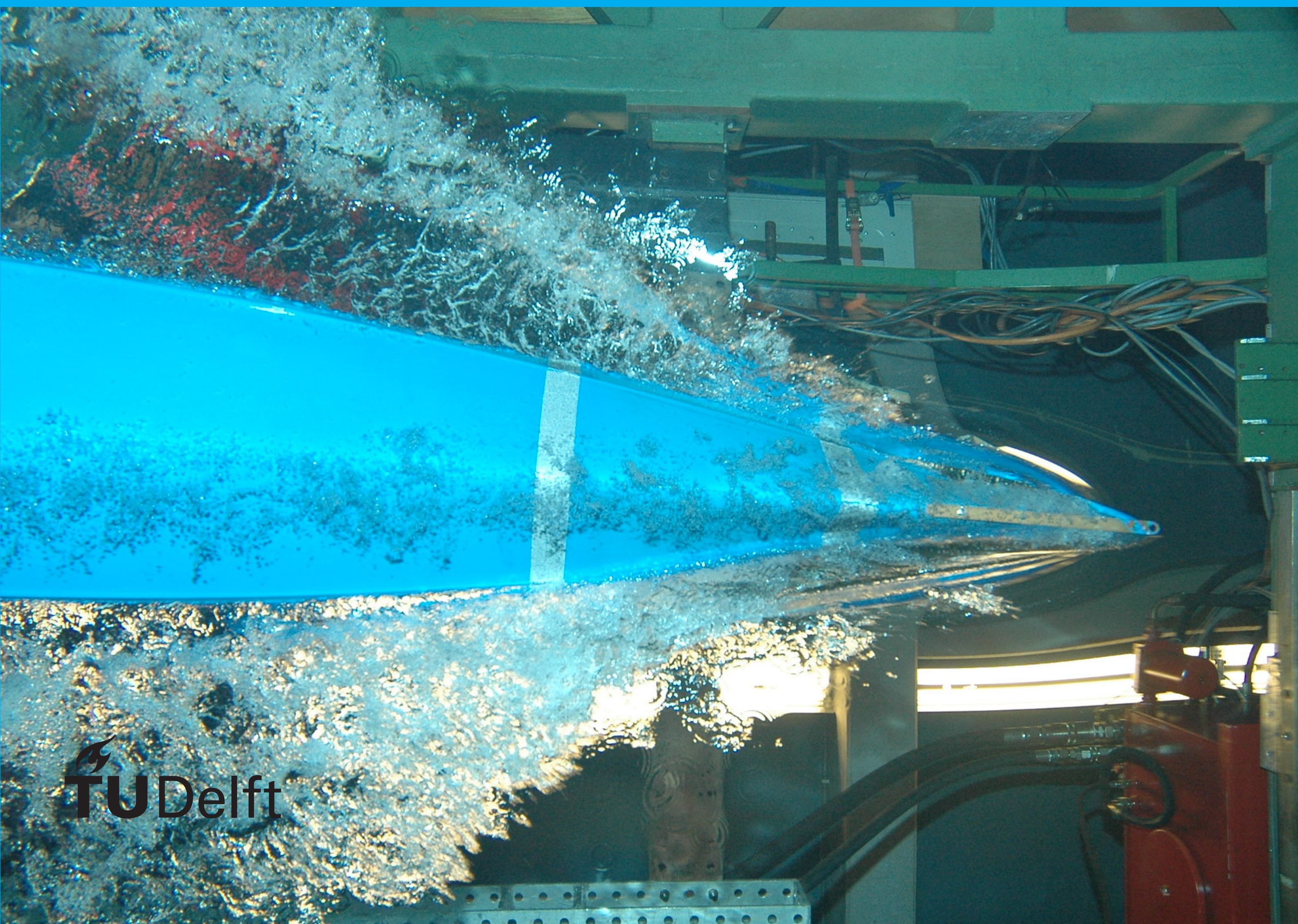
Building a smart rejector for detecting hate speech

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

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, 4]. There are many news articles about events where hate spread on online platforms lead to acts of violence [9, 12–14]. 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 [14].

Furthermore, there is growing interest from governmental institutions in tackling hate speech. Some countries, such as France or Canada, introduced laws to prohibit hate speech [5]. 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 [2]. This Code of Conduct requests companies to prohibit hate speech and report their progress every year [2]. According to the European Commission, hate speech is defined as ”publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin” [2]. The most recent report from 2021 stated that Facebook is most successful in removing hate speech as they claim that they removed 70.2% of all hateful content in 2021 [2]. Twitter is ranked the lowest with 49.5% [2]. However, we should take these numbers with a grain of salt. One article found in internal communication from Facebook that this percentage is much lower, around 3-5% [7].

Manual moderation of hateful content is still the most reliable solution but simply infeasible due to the large amount of content generated by the many users [4]. Therefore, Facebook, for example, adopts both reactive and proactive content moderation [10]. Reactive manual moderation is when users are flagging (also known as reporting) hateful content [10]. The main benefit of this is that many users can check a large amount of content. However, the problem remains that the users are still seeing hateful content for some time. Proactive moderation is either done automatically using detection algorithms or by a dedicated group of employees from Facebook [10]. However, it remains unclear how companies such as Facebook are automatically moderating content since their exact practices are not publicly available [10].

There exist different methods for automatically detecting hateful content when looking at current practices from literature. Most use Machine Learning (ML) algorithms since these tend to be the most promising for their detection performance at a large scale [4, 6]. These algorithms can range from traditional ML methods such as Support Vector Machine or Decision Tree to Deep Learning algorithms [6].

However, these algorithms can be unreliable as they often perform poor on deployment data [4]. For instance, one paper found that most research in hate speech detection overestimates the performance of the automated detection methods [3]. The authors found that there is a significant performance drop when the detection algorithms are trained on one dataset and evaluated on another [3]. Furthermore, internal communication at Facebook indicates that ML algorithms are still not effective enough [7].

Therefore, there is a need for a *socio-technical* or *human-machine co-creation* [16] system that combines the advantages of both humans (cognitive abilities and ability to make judgements) and machines (automation and performance). A system where humans and machines work together to detect hate speech more effectively. This system should be a *machine-assisting-human* system where ML models are helping humans to detect hateful content automatically and where humans can make the final decisions (*human-in-the-loop*) when the model is not confident enough [16]. This leads to our main research question:

RQ1. How can we use human computation in detecting hate speech?

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 [8]. 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. So far, ML with rejection has not been used in hate speech detection yet. Therefore, the goal of this thesis project is to build the first *smart rejector for detecting hate speech*.
But how can we determine whether to reject or accept a prediction? We can argue that there are gains in making correct predictions and costs of rejection and making incorrect predictions. More specifically, we should weigh cost values for False Negative (FN), labelling something as non-hateful when it is, and False Positive (FP) predictions, labelling something as hateful when it is not, according to the task [15]. So, we need a metric that measures the cost-effectiveness of the smart rejector. We can use the resulting metric to determine when to reject/accept predictions by maximizing cost-effectiveness. Therefore, our first sub research question is as follows:

RQ1.1 How can we measure the cost-effectiveness of the reject option?

The idea of ML models with rejection is that predictions are rejected when the confidence of the prediction is too low. However, there also exist cases for which the ML model produces high confidence but incorrect predictions. These high confident errors are also called *unknown unknowns* [11]. We can further improve the smart rejector by detecting these unknown unknowns. This leads to our second sub research question:

RQ1.2 How can we find the unknown unknowns?

Finally, we need to find out how we can combine our findings into one smart rejection system which leads to our final sub research question:

RQ1.3 How can we build the smart rejector?

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. Machine Learning models with rejection

Explain the different architectures of ML with rejection

Explain the different types of confidence metrics

Provide examples of ML models with rejection from other domains

2.2. Hate speech detection

Give some examples of existing hate speech detection methods from literature that for example use traditional Machine Learning algrorithms or Bag of Words

2.3. Unknown unknown detection

Give examples of existing unknown unknown 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-effectiveness metric

Explain the original metric from De Stefano

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

3.3. Unknown unknowns

Explain our method for detecting unknown unknowns

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. Cost values

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

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 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.