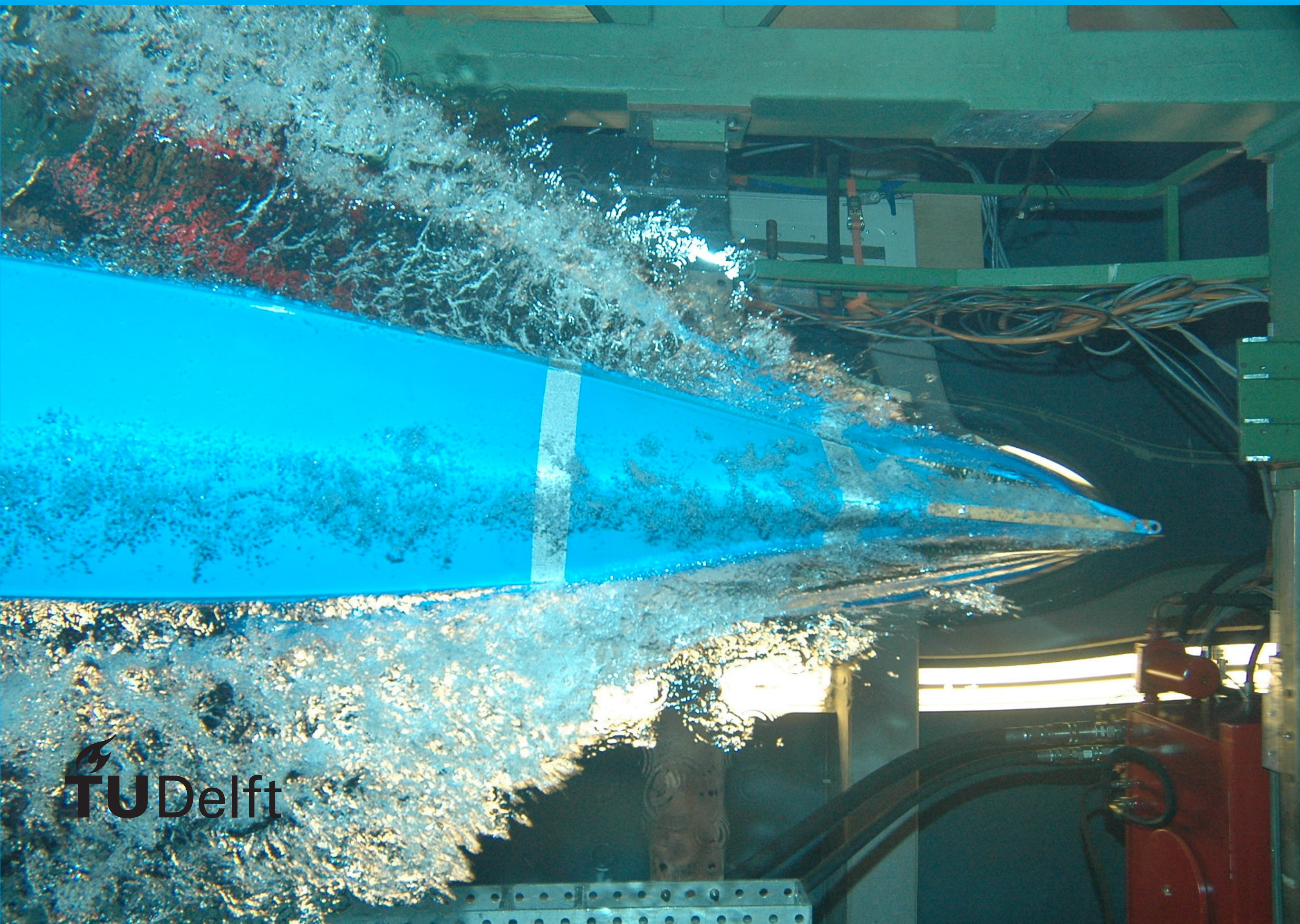
Building a smart rejector for detecting hate speech

Philippe Lammerts



TU Delft

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

Preface

Preface...

Philippe Lammerts
Delft, January 2013

Contents

1 Introduction

The amount of hateful content that is spreaded online on social media platforms remains a serious problem. Ignoring its presence can harm people and might even result into actual acts of violence and other conflicts [1, 4]. There exist many news articles that describe examples of these events where hate spreaded on online platforms lead to actual violence [9, 12–14]. One 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 to tackle the problem of online 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 the companies to prohibit hate speech and to 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 succeful in removing hate speech as they claim to have 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 actually much lower around 3-5% [7].

Manual moderation of hateful content is still the most reliable solution but simply infeasible due the large amount of content that is generated by the many users [4]. Therefore, Facebook adopts both reactive and proactive manual content moderation [10]. Reactive manual moderation is done by Facebook's users by flagging, or reporting, content [10]. The benefit here is that the large amount of data can be checked by many users. But the problem here is that users are still exposed to 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 hidden away from the public [10].

There exist different methods for automatically detecting hateful content when looking at current practices from literature. Most methods use Machine Learning (ML) algorithms since these tend to be the most promising for the their reasonable 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 models can be unreliable as they often perform poor on deployment data [4]. For instance, internal communication at Facebook indicates that ML methods are still not effective enough [7]. Furthermore, one paper found that most research in hate speech detection oversestimate the performance of their detection methods [3]. The authors found that the performance of the detection models drops significantly when they are trained on one dataset and evalauted on another [3].

Therefore, there is a need for a *socio-technical* or *human-machine co-creation* [15] 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 than manual moderation. 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 [15]. This leads to our main research question:

RQ1. How can we use human computing to improve hate speech detection?

Here comes ML models with a reject option into place. The goal of this reject option is to reject a ML prediction when the risk of making an incorrect prediction is too high and to defer the prediction task to a human [8]. This has several advantages. First, the utility of the ML is increased as only the most confident (and possibly the most correct) predictions are accepted. Second, less human effort is required as the machine is handling all predictions tasks where only a small fraction needs to be checked by a human. So far, machine learning with rejection has not been applied to the domain of hate speech detection. Therefore, the goal of this Thesis project is build the first *smart rejection system for detecting hate speech*.

But how can we determine wheher to reject or accept a prediction? We can argue that there gains of making correct predictions and costs for rejecting and making incorrect predictions. We can also argue that the cost of a False Negative (FP) prediction (labeling something as non-hateful while it actually is hateful) is greater than the cost of a False Positive (FP) prediction (labeling something as hateful while it is actually not) in the context of hate speech. The consequences of showing hateful content are worse than hiding neutral content from social media users. Therefore, there is need for a metric that measures the cost-effectiveness of the smart rejector. We can use the resulting metric to find out when to reject/accept predictions by maximizing the cost-effectiveness. This leads to our first sub research question:

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 a high confident but incorrect prediction. 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 these findings into one smart rejection system which leads to our final sub research question:

RQ1.3 How can we build the smart rejector?

ntributions

of thesis report

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] 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.