Building a smart rejector for hate speech detection

Philippe Lammerts
4563182
24-05-2022

Supervised by:
Dr. J. Yang
Dr. Y-C. Hsu
P. Lippmann

TU Delft

Contents

Context

- The amount of hateful content online is a growing concern
- It harms people and can even lead to acts of violence [1-5]
- Tackling hate speech is in the interest of all:
  - Governments
  - Social media companies
  - Everybody

The New York Times

As Officials Look Away, Hate Speech in India Nears Dangerous Levels

The New York Times

A Genocide Incited on Facebook, With Posts From Myanmar's Military

TU Delft

[1] 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.

[2] 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.

[3] https://www.nytimes.com/2022/02/08/world/asia/india-hate-speech-muslims.html.

[4] https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

[5] https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence

Context

Two ways to detect hate speech [1]:

1. **Reactive moderation**
   a. Flagging/reporting
2. **Proactive moderation**
   a. Manually
   b. Automatically

TU Delft

[1] Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. Harv. L. Rev., 131, 1598.

Problem

Manual proactive moderation

Human moderators

+ Most reliable
- Infeasible

Automated proactive moderation

Machine Learning algorithms

+ Fast
- Can be unreliable [1, 2]:
  - Performs poor on deployment data
  - 69% F1-score drop when using different test datasets [2]

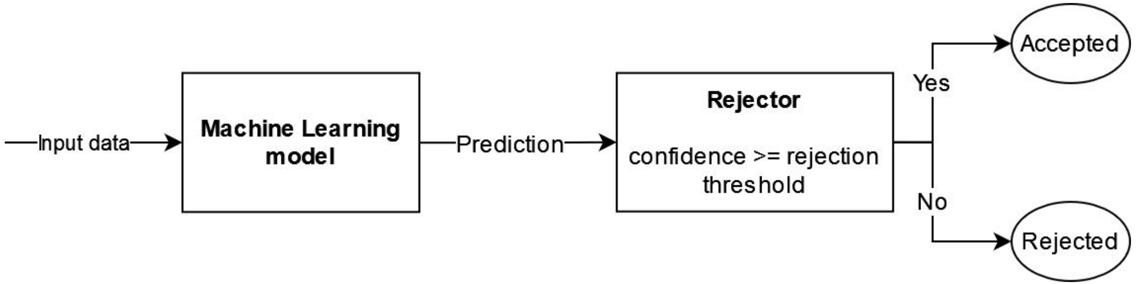
TU Delft

[1] 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.

[2] Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is" love" evading hate speech detection. In Proceedings of the 11th ACM workshop on artificial intelligence and security (pp. 2-12).

Rejection

“The goal of a machine learning model's reject option is to abstain from making a prediction when a model receives a test example where the risk of making a misprediction is too large.” [1]



[1] Hendrickx, Kilian, et al. "Machine Learning with a Reject Option: A survey." *arXiv preprint arXiv:2107.11277* (2021).

Research questions

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

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 relative costs of rejections and True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions?

SRQ2 Can unknown (un)known detection further improve the reject option?

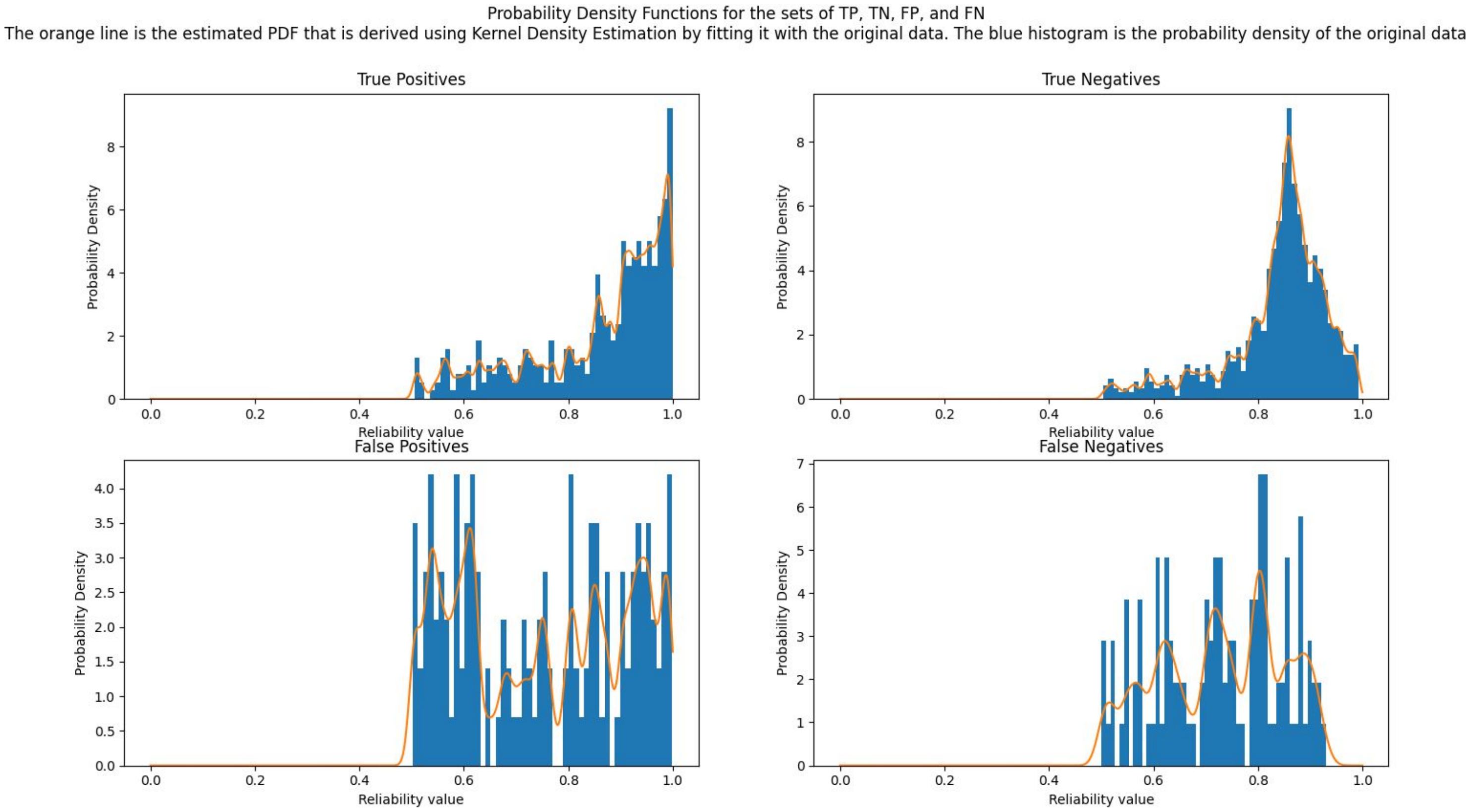
Part 1: rejection metric

SRQ1.1 How can we measure the utility of Machine Learning models with a reject option?

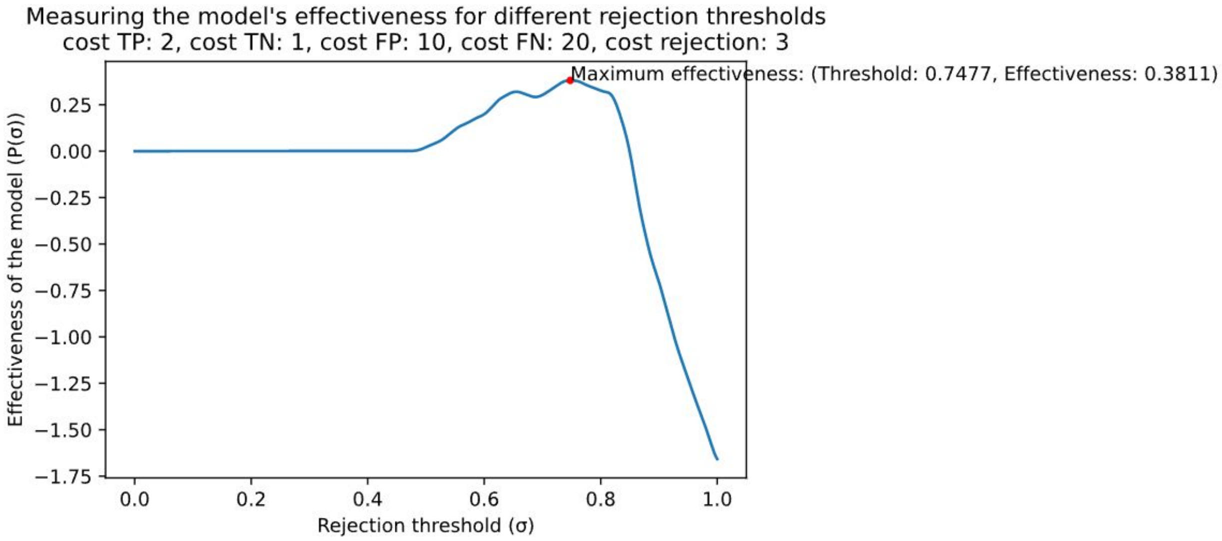
- Metric that measures the effectiveness of the reject option
- Advanced version of the metric in [1]
- Calculation based on:
  - Gain of True Positive
  - Gain of True Negative
  - Cost of False Positive
  - Cost of False Negative
  - Cost of rejection
  - List of predictions with confidence values

[1] 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. https://doi.org/10.1109/5326.827457

Part 1: rejection metric - example



Part 1: rejection metric - example



TU Delft

Part 2: costs of predictions

SRQ1.2 How can we determine the relative costs of rejections and True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions?

- Objective cost analysis in hate speech is infeasible [1]
- We focus on subjective cost analysis
- Magnitude Estimation scale
  - Unbounded rating scale
  - Provides ratio data
  - Used in many different domains [2-6]

[1] Sunstein, C. R. (2018). Does the Clear and Present Danger Test Survive Cost-Benefit Analysis?. Cornell L. Rev., 104, 1775.

[2] Maddalena, E., Mizzaro, S., Scholer, F., & Turpin, A. (2017). On crowdsourcing relevance magnitudes for information retrieval evaluation. ACM Transactions on Information Systems (TOIS), 35(3), 1-32.

[3] Lodge, M., & Tursky, B. (1979). Comparisons between category and magnitude scaling of political opinion employing SRC/CPS items. American Political Science Review, 73(1), 50-66.

[4] Lodge, M., Tanenhaus, J., Cross, D., Tursky, B., Foley, M. A., & Foley, H. (1976). The calibration and cross-modal validation of ratio scales of political opinion in survey research. Social Science Research, 5(4), 325-347.

[5] McGee, M. (2004, April). Master usability scaling: magnitude estimation and master scaling applied to usability measurement. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 335-342).

[6] Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. Language, 32-68.

TU Delft

Part 2: costs of predictions

Experiment

Present TP, TN, FP, FN, and rejection scenarios to a group of subjects:

- Show (non)hateful tweet
- Show decision of platform
- Ask subjects whether they (dis)agree with the decision using the ME scale

Progress

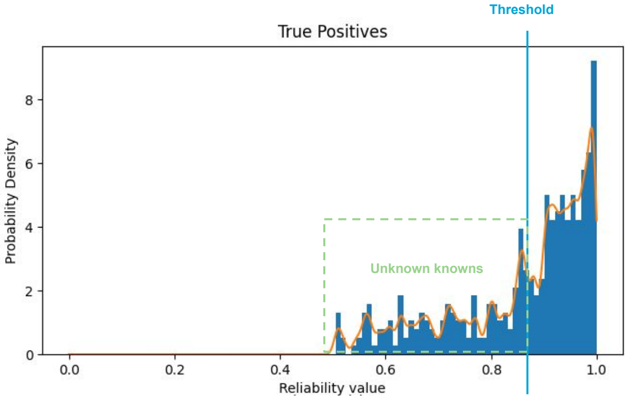
- Pre-registration report is ready
  - Experimental setup
  - Procedure
  - Analysis
- Waiting for approval of Human Research Ethics

TU Delft

Part 3: unknown (un)knowns

SRQ2 Can unknown (un)known detection further improve the reject option?

- A single optimal rejection threshold is not enough
- Different methods for unknown (un)known detection [1-3]



[1] Liu, A., Guerra, S., Fung, I., Matute, G., Kamar, E., & Lasecki, W. (2020, April). Towards hybrid human-AI workflows for unknown unknown detection. In Proceedings of The Web Conference 2020 (pp. 2432-2442).

[2] Attenberg, J., Ipeirotis, P., & Provost, F. (2015). Beat the machine: Challenging humans to find a predictive model's “unknown unknowns”. Journal of Data and Information Quality (JDIQ), 6(1), 1-17.

[3] Lakkaraju, H., Kamar, E., Caruana, R., & Horvitz, E. (2017, February). Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In Thirty-first aaai conference on artificial intelligence.

Planning

- Currently finishing part 2
- Combine findings of parts 1 + 2 and create a paper submission for:
    - *The 10th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022)*
    - Deadline June 24, 2022
- Work on part 3 after the deadline
- Finish around the end of Q1 2022-2023