Oppositeness-based Hate Speech Detection on Social-Media Platforms

Dinuja Perera

content

- **Introduction**
- **Recap**
- **Transformer Models**
- **Ongoing Work**

Introduction

What is hate speech?

We define hate speech as a direct attack on people based on what we call protected characteristics—**race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability**. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. [1]

- Facebook –

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of **race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease**. [2]

- Twitter -

"We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: **age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation,** victims of a major violent event and their kin, and veteran Status“[3]

- YouTube -

[1] “Community standards.” [Online]. Available: https://www.facebook.com/ communitystandards/objectionable_content/

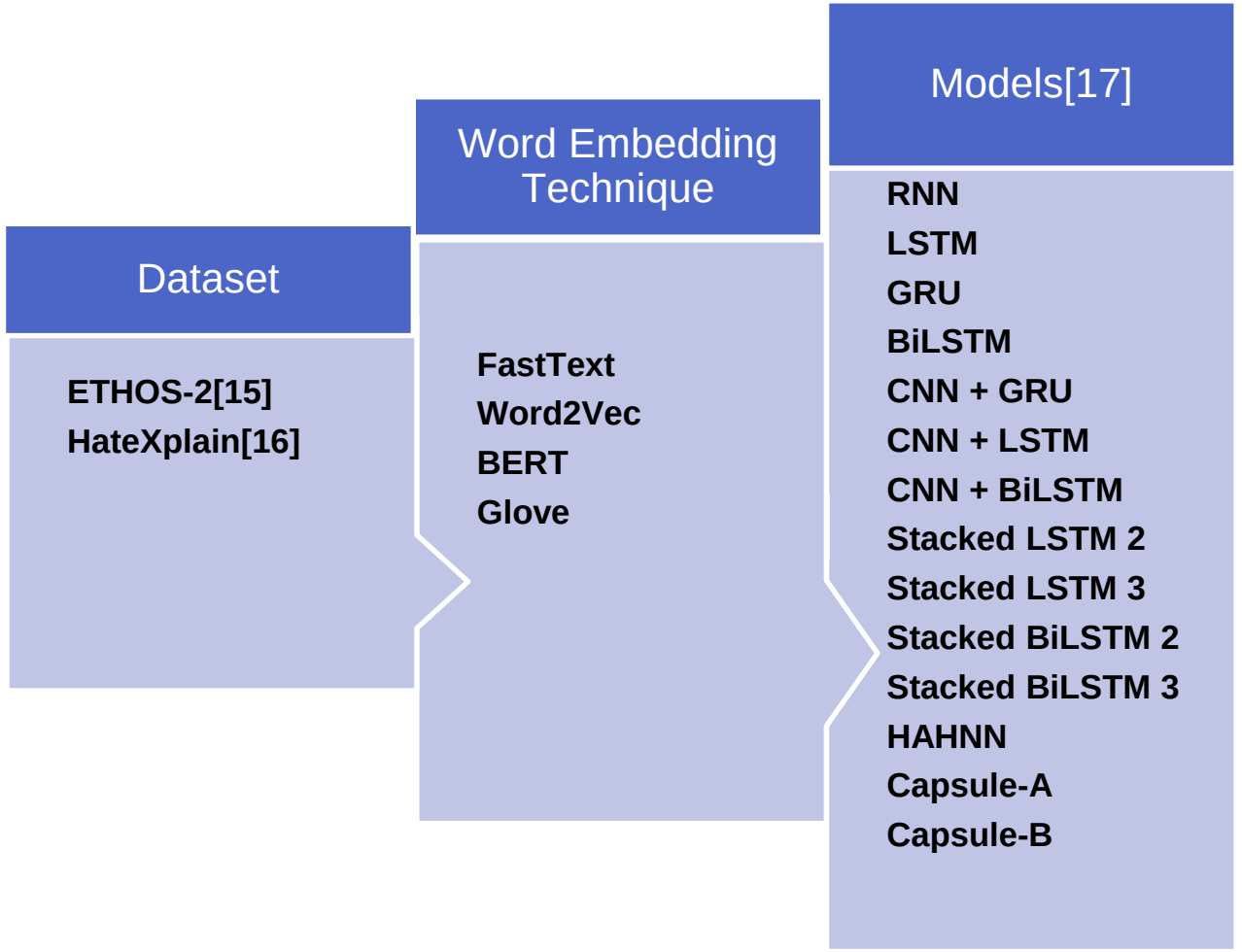
[2] “twitters policy on hate help.” [Online]. Available: https://archive.org/details/perma_cc_2XYS-VWJB

[3] “Hate speech policy - ful conduct | twitter youtube help.” [Online]. Available: https://support.google. com/youtube/answer/2801939?hl$=$en

INTRODUCTION

- Challenge of detecting hate speech within online user communication due to its vast scope and the complexity
- Secure the freedom of speech
- Novel approach in HS detection using oppositeness measure

Recap



[4]vMollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, “Ethos: an online hate speech detection dataset,” arXiv preprint arXiv:2006.08328, 2020
[5] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “Hatexplain: A benchmark dataset for explainable hate speech detection,” arXiv preprint arXiv:2012.10289, 2020
[6]Senevirathne, L., Demotte, P., Karunanayake, B., Munasinghe, U., Ranathunga, S.: Sentiment analysis for sinhala language using deep learning techniques. arXiv preprint arXiv:2011.07280 (2020)

2-ETHOS dataset

- ETHOS Binary Dataset
- onlinE haTe speecH detectiOn dataSet (ETHOS) dataset
- Mollas et al. (2020)
- Comments from YouTube and Reddit.

  YouTube data → Hatebusters [7]

  Reddit data → Public Reddit Data Repository [8]

• The classification is done by contributors to a crowd-sourcing platform.

• The dataset has two variants: binary and multi-label

• ETHOS-2 → hate or non-hate based

• ETHOS-multi → violence, gender, race, ability, religion, and sexual orientation.

• Dataset contain typos, misspelling, and offensive content

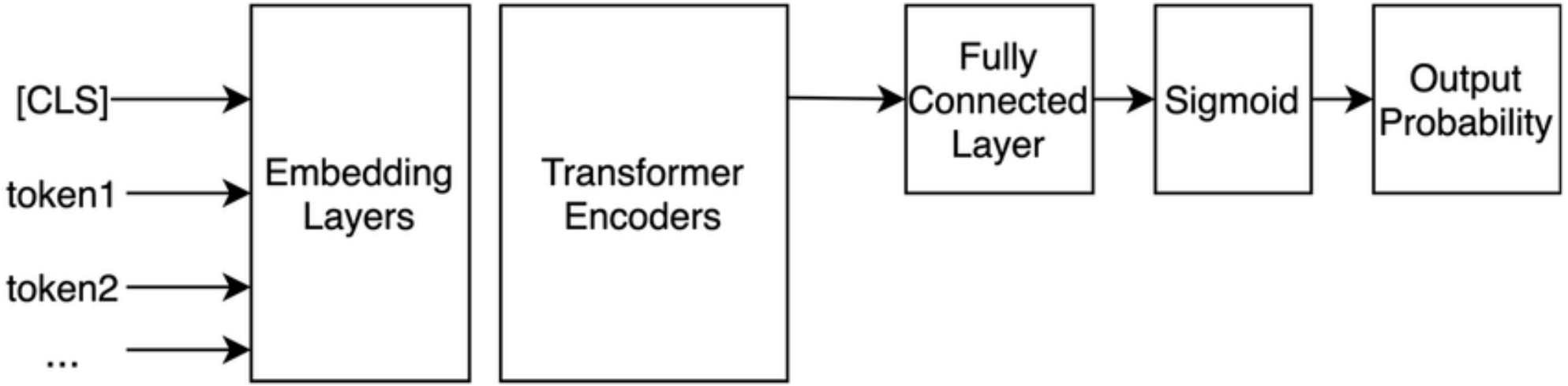
[7] Anagnostou, A., Mollas, I., and Tsoumakas, G. (2018). Hatebusters: A Web Application for Actively Reporting YouTube Hate Speech. In IJCAI, pages 5796–5798.

[8] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In Proceedings of the International AAAI Conference on Web and Social Media, volume 14, pages 830–839

BART large mnli model

- By Facebook
- Yin et al.
- Pre-trained NLI models as a ready-made zero-shot sequence classifiers .
- Posing the sequence to be classified as the NLI premise and to construct a hypothesis from each candidate label .
- The probabilities for entailment and contradictions convertes to label probabilities .

xlm-roberta-large-xnli



[7] Xu, Jian & Zhai, Yuqing. (2021). A Toxic Comment Classification Model Based on Ensemble. Journal of Physics: Conference Series. 1873. 012080. 10.1088/1742-6596/1873/1/012080

Results

- **BART large mnli model**

```
You should know women's sports are a joke {'labels': ['Hate', 'Not hate'], 'scores': [0.5838798880577087,
0.4161201417446136 5]}

who gives aa fuck about shit diamonds from the jews fuck them both {'labels': ['Hate', 'Not hate'],
'scores': [0.6683055758476257, 0.3316943943500519]}

down their own neighborhoods when a black thug gets killed by law enforcement? Must be because of low IQ.
{'labels': ['Hate', 'Not hate'], 'scores': [0.8895377516746521, 0.1104622632265091]}
```

- **xlm-roberta-large-xnli**

```
You should know women's sports are a joke {'labels': ['Hate', 'Not hate'], 'scores': [0.7833456993103027,
0.2166542708873748 8]}

who gives aa fuck about shit diamonds from the jews fuck them both {'labels': ['Hate', 'Not hate'],
'scores': [0.7660819888114929, 0.2339180111885070 8]}

Why do negroes tear down their own neighborhoods when a black thug gets killed by law enforcement? Must be
because of low IQ. {'labels': ['Hate', 'Not hate'], 'scores': [0.7356250286102295, 0.2643749713897705]}
```

Results

Model	Comment	Hate score
BART large mnli	You should know women's sports are a joke	0.584
	who gives aa fuck about shit diamonds from the jews fuck them both	0.668
	down their own neighborhoods when a black thug gets killed by law enforcement? Must be because of low IQ.	0.890
xlm-roberta-large-xnli	You should know women's sports are a joke	0.783
	who gives aa fuck about shit diamonds from the jews fuck them both	0.766
	down their own neighborhoods when a black thug gets killed by law enforcement? Must be because of low IQ.	0.736

REFERENCES

[1] “Community standards.” [Online]. Available: https://www.facebook.com/ communitystandards/objectionable_content/

[2] “twitters policy on hate help.” [Online]. Available: https://archive.org/details/perma_cc_2XYS-VWJB

[3] “Hate speech policy - ful conduct | twitter youtube help.” [Online]. Available: https://support.google.com/youtube/answer/2801939?hl$=$en

[4]vMollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, “Ethos: an online hate speech detection dataset,” arXiv preprint arXiv:2006.08328, 2020

[5] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “Hatexplain: A benchmark dataset for explainable hate speech detection,” arXiv preprint arXiv:2012.10289, 2020

[6]Senevirathne, L., Demotte, P., Karunanayake, B., Munasinghe, U., Ranathunga, S.: Sentiment analysis for sinhala language using deep learning techniques. arXiv preprint arXiv:2011.07280 (2020)

[7] Xu, Jian & Zhai, Yuqing. (2021). A Toxic Comment Classification Model Based on Ensemble. Journal of Physics: Conference Series. 1873. 012080. 10.1088/1742-6596/1873/1/012080.

Thank you!

Any Questions?

You can find me at:
dinuja.21@cse.mrt.ac.lk