Hate Speech Detection using Deep Neural Networks

Dinuja Perera

Overview

INTRODUCTION

RESEARCH PROBLEM

CHOOSING THE DATASET

ON-GOING WORK

RESULTS

INTRODUCTION

- Growth of social-media usage, raises a platform to a new kind of social dilemma namely cyberbullying

  - Duke and Duchess of Sussex has been targeted intentionally via many anti-Meghan and Harry accounts solely created to spread hateful content on Twitter [9]

- HS detection is still a challenge for the research community and policy makers as humans find loopholes to trick those algorithms [8]

- Challenge of detecting hate speech within online user communication due to its vast scope and the complexity

[4] A. Oboler, “Solving antisemitic hate speech in social media through a global approach to local action,” in Volume 5 Confronting Antisemitism in Modern Media, the Legal and Political Worlds. De Gruyter, 2021, pp. 343–368. 2
[5] “Twitter hate accounts targeting meghan and harry, duke and duchess of sussex,” Oct 2021. [Online]. Available: https://www.msn.com/en-us/news/politics/ organized-campaign-targeted-harry-and-meghan-on-twitter-report/ar-AAQ180P
[6] T. M. Massaro, “Equality and freedom of expression: The hate speech dilemma,” Wm. & Mary L. Rev., vol. 32, p. 211, 1990. 4

Data sets

Dataset	Size	Citations	Year	Classes	Description
ETHOS (Binary)	998	17	2020	Hate speech, Not hate speech	Generated from YouTube and Reddit comments
ETHOS (Multi)	433	9	2020	Gender, Race, Violence, national-origin directed-vs generalized,Religion, Disability, sexual-orientation	Generated from YouTube and Reddit comments
Twitter15	1381	257	2015	True Rumours, False Rumours, Unverified Rumours, Non-Rumours	Crawled from Twitter
Twitter16	1181	389	2016	True Rumours, False Rumours, Unverified Rumours, Non-Rumours	Crawled from Twitter
HateXplain	20K	30	2021	hateful, offensive, normal	Posts from Twitter2 and Gab3,in addition to classifying, target communities of the post are anotated by mazon Mechanical Turk workers.
ElSherief et al.	27,330	94	2019	Archaic Class Disability Ethnicity Gender Nationality Religion SexOrient	Twitter Streaming API, 1% Collection of daily tweets posted for 18 month with the information on users who posted them
de Gibert et al	9916	145	2018	Hate, Relation, Not-Hate	Crawled from Storm-front platform and contains 11% of hate
Gao and Huang	1528	115	2017	Binary (Hate / not)	Crawled from Fox News platform and contains 28% of hate.
Ribeiro et al. [	4972	137	2018	Binary (Hate, Not-hate)	Crawled from Twitter platform and contains 11% of hate
Waseem and Hovy	16914	901	2016	Sexist, Racist, Not	Dataset from Twitter platform that contains 32% hate

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
- Hate speech detection using static BERT embeddings (2021)

- Analyze the performance of hate speech detection classifier by replacing or integrating the word embeddings (fastText (FT), GloVe (GV) or FT + GV) with static BERT embeddings (BE)

- Detecting Hate Speech with GPT-3 (2022)

- OpenAI's GPT-3 , generate hateful text

[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

HateXplain data set

- A Benchmark Dataset for Explainable Hate Speech Detection
- Each post is annotated from three different perspectives:

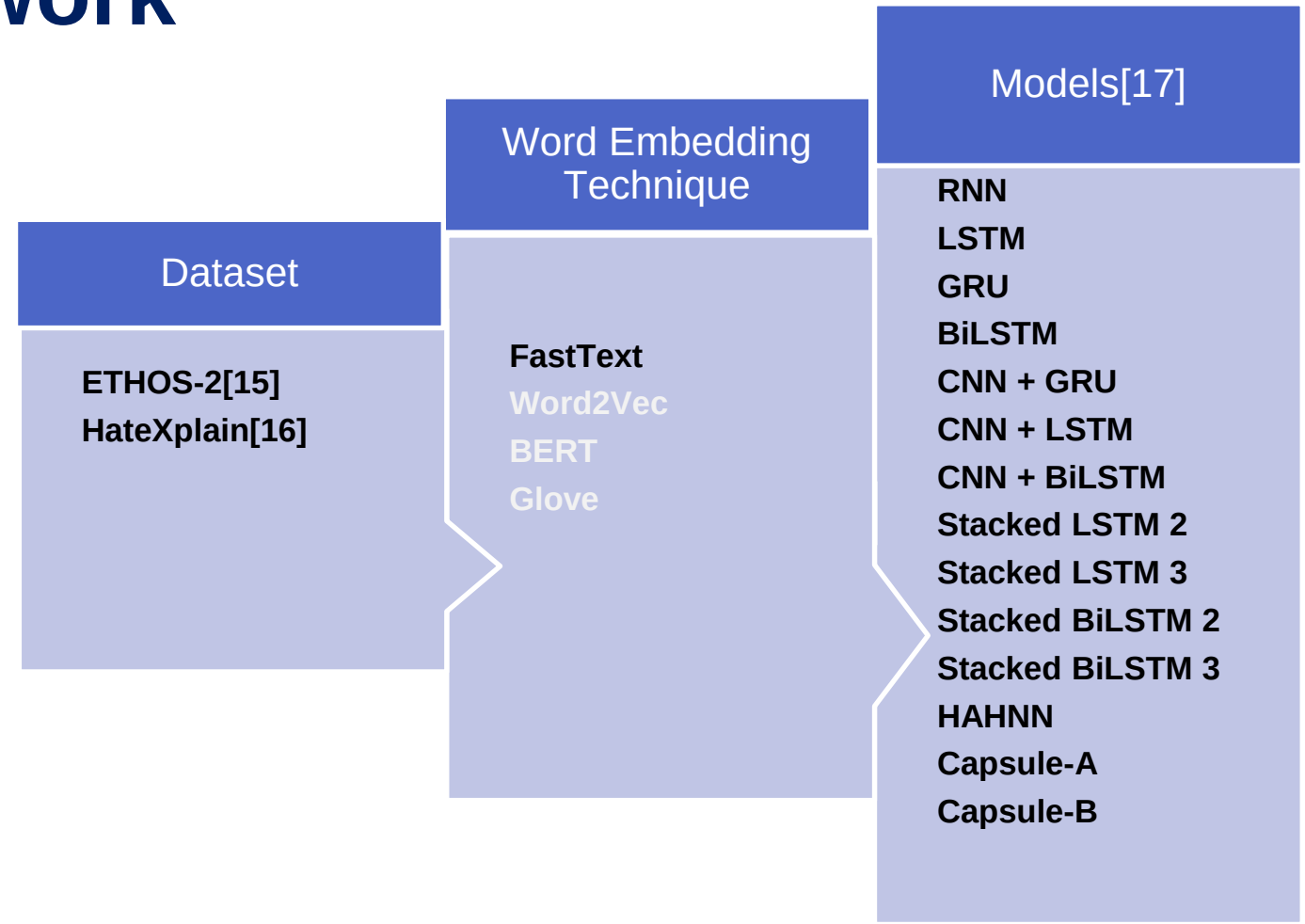
  3-class classification (hate, offensive or normal),

  the target community

  the rationales → the portions of the post on which their labelling decision (as hate, offensive or normal)

Model [Token Method]	**Performance**		
	Acc.↑	Macro F1↑	AUROC↑
CNN-GRU [LIME]	0.627	0.606	0.793
BiRNN [LIME]	0.595	0.575	0.767
BiRNN-Attn [Attn]	0.621	0.614	0.795
BiRNN-Attn [LIME]	0.621	0.614	0.795
BiRNN-**HateXplain** [Attn]	0.629	0.629	0.805
BiRNN-**HateXplain** [LIME]	0.629	0.629	0.805
BERT [Attn]	0.690	0.674	0.843
BERT [LIME]	0.690	0.674	0.843
BERT-**HateXplain** [Attn]	**0.698**	**0.687**	**0.851**
BERT-**HateXplain** [LIME]	**0.698**	**0.687**	**0.851**

On-going work



[1]vMollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, “Ethos: an online hate speech detection dataset,” arXiv preprint arXiv:2006.08328, 2020
[2] 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
[3]]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)

Results

Model	ETHOS-2		HateXplain	
	FastText		FatText	
	Accuracy	F1-score	Accuracy	F1-score
RNN	0.563	0.582		
LSTM	0.584	0.641		
GRU	0.558	0.624		
BiLSTM	0.628	0.643		
CNN+GRU	0.622	0.632		
CNN + LSTM	0.680	0.688		
Stacked LSTM 3	0.566	0.577		
CNN + BiLSTM				
Stacked LSTM2				
HAHNN				
Capsule-A				
Capsule-B				

REFERENCES

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

[2] 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

[3]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)

[4] A. Oboler, “Solving antisemitic hate speech in social media through a global approach to local action,” in Volume 5 Confronting Antisemitism in Modern Media, the Legal and Political Worlds. De Gruyter, 2021, pp. 343–368. 2

[5] “Twitter hate accounts targeting meghan and harry, duke and duchess of sussex,” Oct 2021. [Online]. Available: https://www.msn.com/en-us/news/politics/ organized-campaign-targeted-harry-and-meghan-on-twitter-report/ar-AAQ180P

[6] T. M. Massaro, “Equality and freedom of expression: The hate speech dilemma,” Wm. & Mary L. Rev., vol. 32, p. 211, 1990. 4

[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

Thank you!

Any Questions?

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