Oppositeness-based Hate Speech Detection on Social-Media Platforms

Dinuja Perera

content

- Prelude
- Background
- Current Methods
- Future Research Direction

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

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]

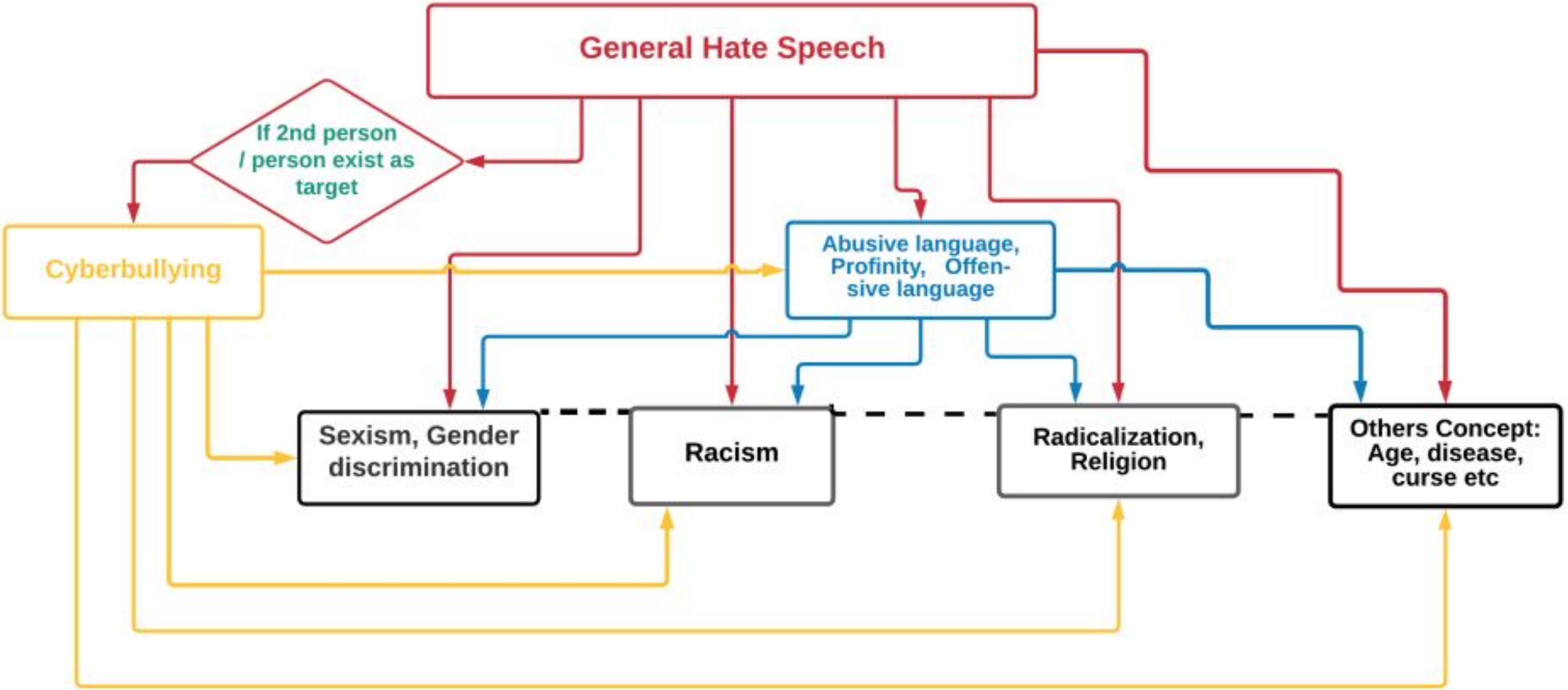
- YouTube -

[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

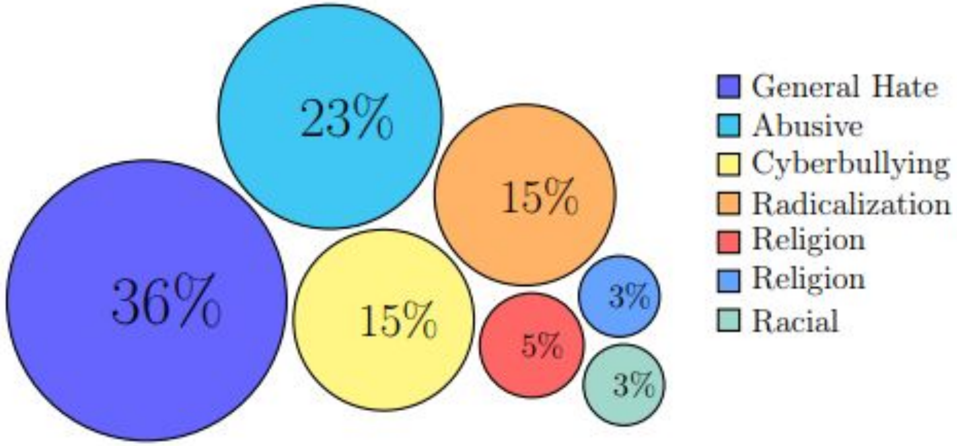
[3] “Hate speech policy - ful conduct | twitter youtube help.” [Online]. Available: https://support.google. com/youtube/answer/2801939?hl$=$en

Introduction cont.

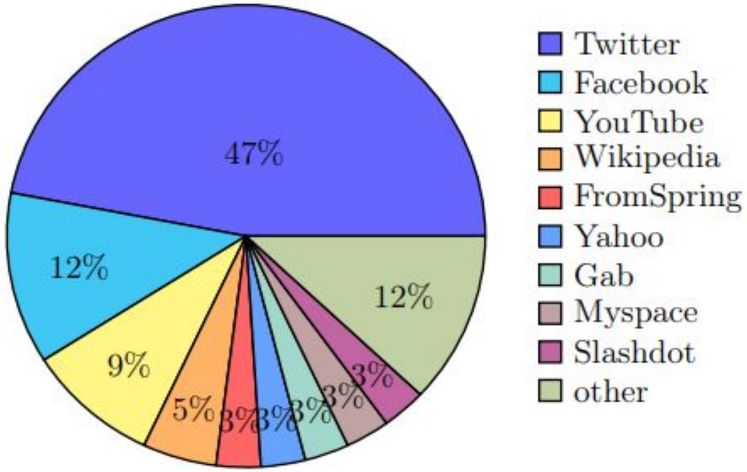

General Hate Speech
If 2nd person / person exist as target
Cyberbullying
Abusive language, Profinity, Offen-sive language
Sexism, Gender discrimination
Racism
Radicalization, Religion
Others Concept: Age, disease, curse etc

Background Study

- Jahan and M. Oussalah
- systematic literature review from Google scholar and ACM digital library databases for all documents related to hate speech published between 2000 and 2021[4].
- 463 articles

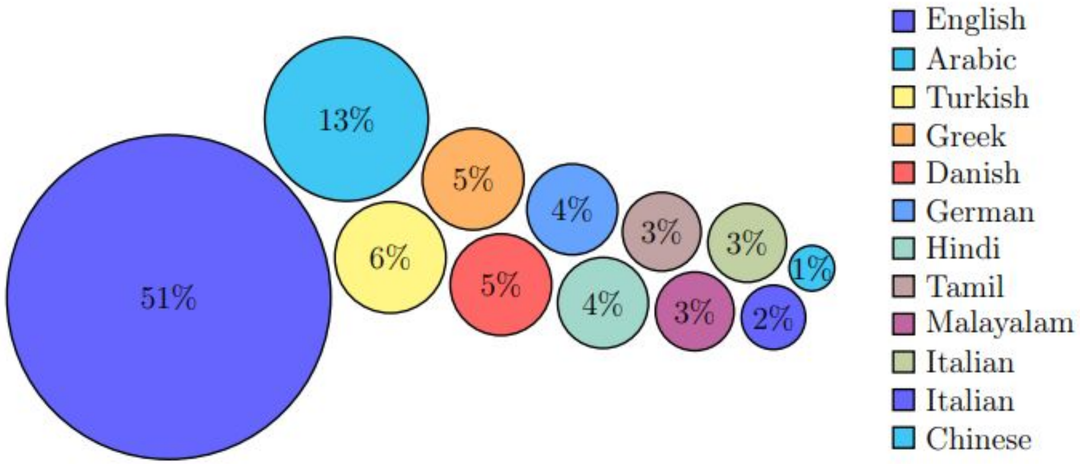


Percentage of Hate Speech categories

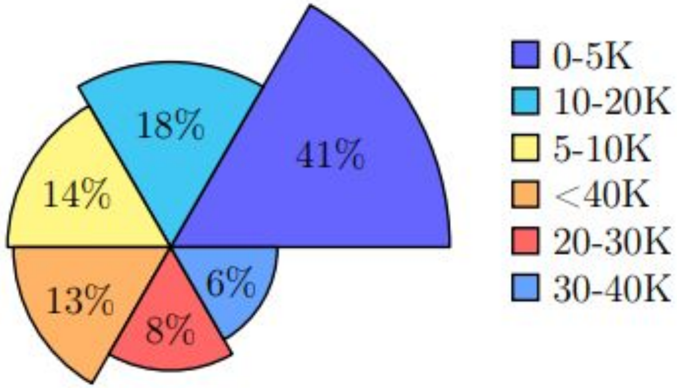


Percentage available data collection platforms

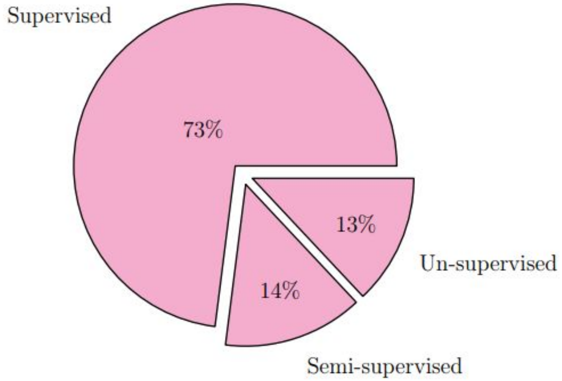
[4] M. S. Jahan and M. Oussalah, “A systematic review of hate speech automatic detection using natural language processing,” arXiv preprint arXiv:2106.00742, 2021.



Statistics of HS work in various languages



Statistics of the size of the dataset



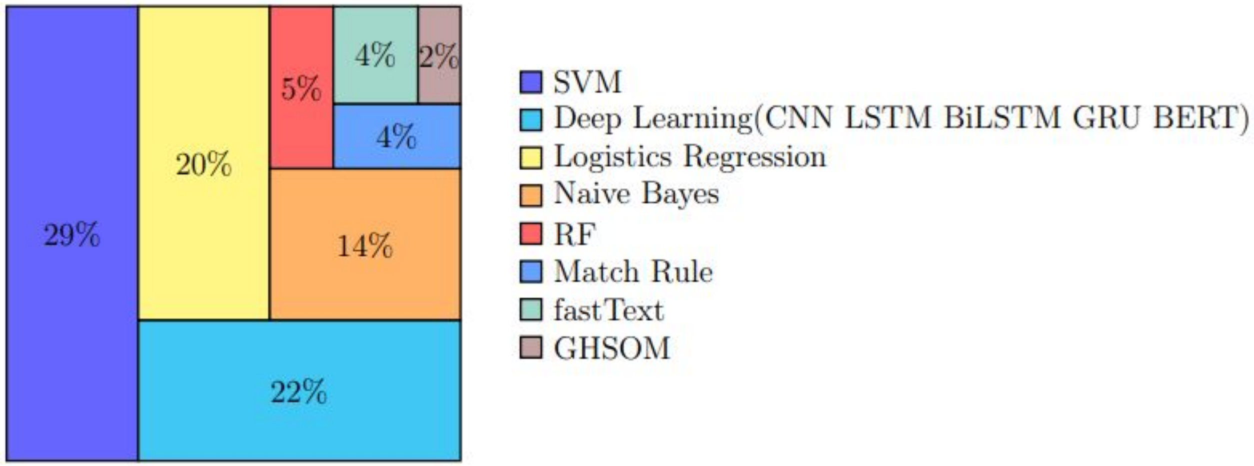
Percentage of Types of Machine Learning approach

Detecting offensive language in social media to protect adolescent online safety,
- Unsupervised method and lexical & syntactic features to achieve 98% accuracy[5].
- Supervised and semi-supervised methods that have shown close or better performance [6] [7].

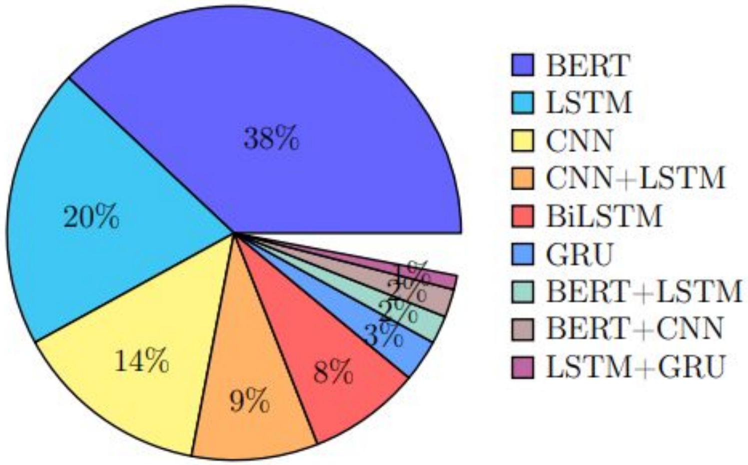
[5] Chen, Y., Zhou, Y., Zhu, S., Xu, H., 2012. Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE.

[6] Abozinadah, E.A., 2016. Improved micro-blog classification for detecting abusive arabic twitter accounts. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol 6.

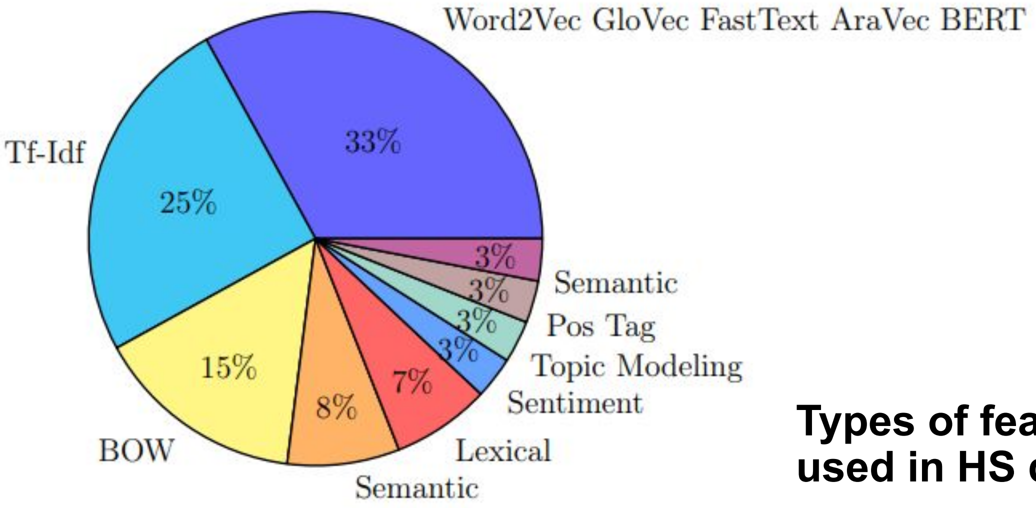
[7] Badjatiya, P., Gupta, S., Gupta, M., Varma, V., 2017. Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion



Types of Supervised and Deep Learning algorithms



Types Deep Learning algorithms



Types of feature embedding techniques used in HS detection

Expectations & challenges

OPEN-SOURCE PLATFORMS OR ALGORITHMS

LANGUAGE AND SYSTEM BARRIERS

CHOOSING THE DATASET

COMPARATIVE STUDIES

MULTILINGUAL RESEARCH

Existing Methods

- Rumor Detection based on oppositeness measure[8]
- Sentiment Analysis for Sinhala Language using Deep Learning Techniques[9]
- Hate Speech Detection with Comment Embeddings[10]
- Hate speech detection using static BERT Embeddings [11]
- Data expansion using back translation and paraphrasing for hate speech detection [12]
- Detecting Hate Speech with GPT-3 [13]

UNDERSTANDING THE OPPOSITENESS MEASURE[14]

- Semantic similarity □ the degree to which extent any two concepts/words/sentences are similar to each other in a given domain

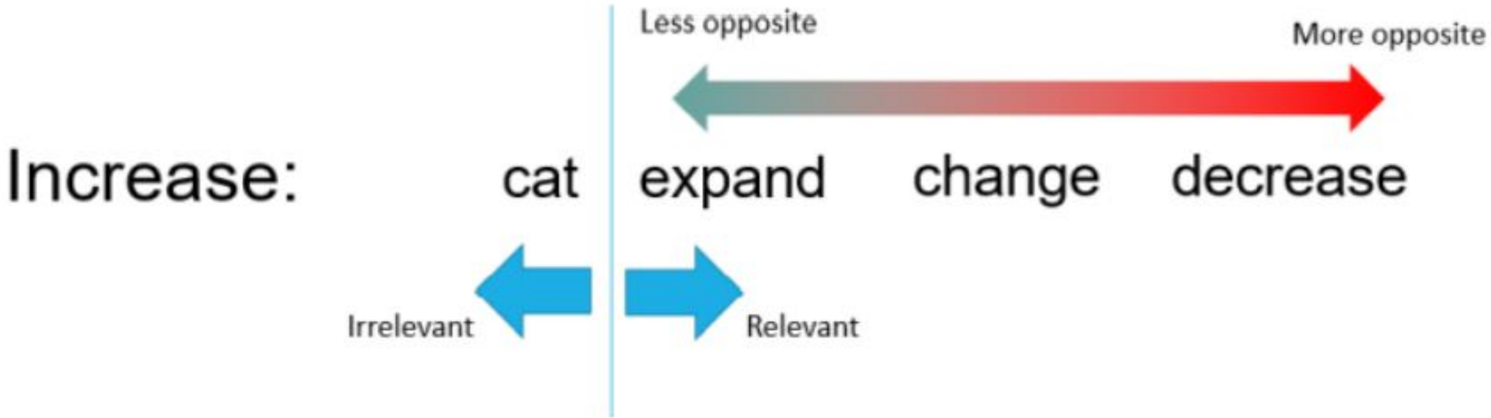
- Complement(semantic similarity) = semantic oppositeness

Word 1	Word2	similarity of two words by Wu and Palmer (1994)	1- similarity
Increase	Cat	0.11	0.89
Increase	Decrease	0.33	0.67
Increase	Change	0.36	0.64
Increase	Expand	0.75	0.25

[14[N. H. N. D. de Silva, “Semantic oppositeness for inconsistency and disagreement detection in natural language,” Ph.D. dissertation, University of Oregon, 2020.

UNDERSTANDING THE OPPOSITENESS MEASURE

Expected optimal word order of expand, decrease, change, and cat in respect to the word increase



- Complement(semantic similarity) != semantic oppositeness.

- w1, w2 --> Words
- L1, L2 --> lemmas of W1 and W2
- a1,a2 --> antonym list of L1 and L2
- n,m --> number of items in a1 and a2

dif 1 = max(sim(L2, a1(1)), sim(L2, a1(2)), ..., sim(L2, a1(n)))

dif 2 = max(sim(L1, a2(1)), sim(L1, a2(2)), ..., sim(L1, a2(m)))

dif = ((dif1/c1)+(dif2/c1))/2

Data sets

Dataset	Size	Citations	Year	Classes	Description
ETHOS (Binary)	998	9	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

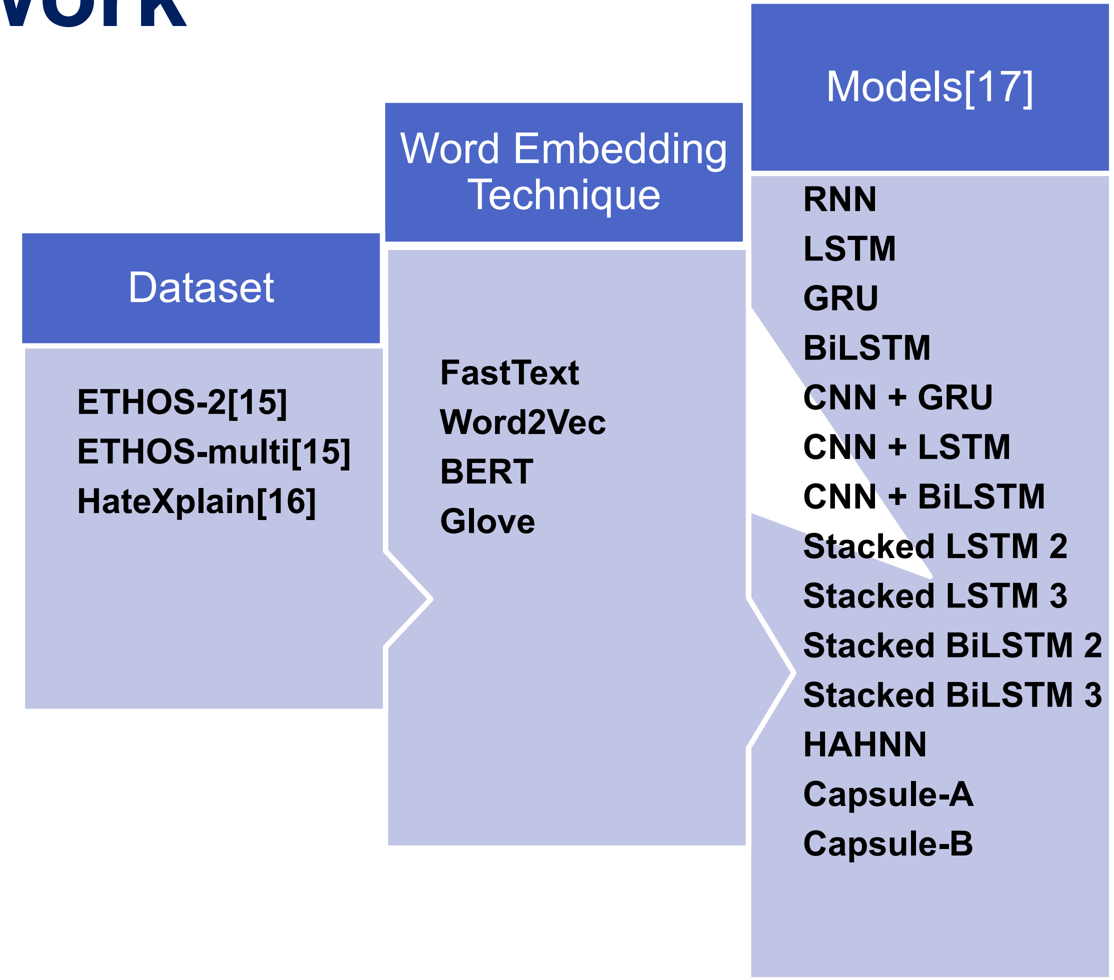
Proposed methodologies

- Function to calculate the oppositeness measure

- Need to crowd relevant datasets that consists user feed-backs along with the main content
- e.g: Twitter15 and Twitter16

- Need to prove that the oppositeness from the same community in the same time frame assist to handle the challenges of detecting HS in scenarios where not raise an opposite response but instead raise approval

On-going work



Dataset

ETHOS-2[15]
ETHOS-multi[15]
HateXplain[16]

Word Embedding Technique

FastText
Word2Vec
BERT
Glove

Models[17]

RNN
LSTM
GRU
BiLSTM
CNN + GRU
CNN + LSTM
CNN + BiLSTM
Stacked LSTM 2
Stacked LSTM 3
Stacked BiLSTM 2
Stacked BiLSTM 3
HAHNN
Capsule-A
Capsule-B

[15]vMollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, “Ethos: an online hate speech detection dataset,” arXiv preprint arXiv:2006.08328, 2020
[16] 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
[17]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)

REFERENCES

[8] N. de Silva and D. Dou, “Semantic oppositeness assisted deep contextual modeling for automatic rumor detection in social networks,” in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 405–415.

[9] 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)

[10] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in Proceedings of the 24th international conference on world wide web, 2015.

[11] G. Rajput, S. K. Sonbhadra, S. Agarwal et al., “Hate speech detection using static bert embeddings,” arXiv preprint arXiv:2106.15537, 2021.

[12] D. Romaissa Beddiar, M. Saroar Jahan, and M. Oussalah, “Data expansion using back translation and paraphrasing for hate speech detection,” arXiv e-prints, pp. arXiv–2106, 2021.

[13] K.-L. Chiu and R. Alexander, “Detecting hate speech with gpt-3,” arXiv preprint arXiv:2103.12407, 2021.

Thank you!

Any Questions?

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