Grenze International Journal of Engineering and Technology, Jul Issue

A Review on Hate Speech Detection using Natural Language Processing

Aparna Susan Varughese[1], Fathima Swaliha T M[2], Nargis Irfan[3] and Prof. Rehannara Beegum T[4]
[1-4]TKM College of Engineering, Kollam, Kerala, India
Email: {aparnavarghese11, swalitm, nargisirfan5}@gmail.com, rehannara.beegum@tkmce.ac.in

***Abstract*—With increased use of social media among various users, there has been a prominent surge in hate speech content in the online platforms. This paper provides a review on how natural language processing is being used to combat this surge. The paper discusses the resources like tools and datasets that are currently available. The features and various methods of detection are explored. The most targeted groups of people are also identified. The languages which have been covered under hate speech detection are also discussed. The challenges that are faced are reviewed.**

***Index Terms*— hate speech, natural language processing, classifiers, languages.**

I. INTRODUCTION

Recent times have seen an increase in hate speech across various social media platforms. This can result in many users being prone to a negative influence online and spoil their experience of using social media by projecting them to an unsafe environment. It has been found that in 2011, social media platforms are used regularly by 70% of the teens, and 19% of them report that nasty and humiliating things about them have been posted online by other users. [1]. Such incidents can influence their behavioural characteristics. Also, such unsafe environments may lead to an increase in crimes and incidents of self-harm. Hence it is the role of social media platforms to take steps to mitigate this.

But this poses a major challenge as content posted on social media platforms is massive, and hence it is not possible to perform a manual filtering of hate messages. This raises the need to develop automatic filtering techniques which can be implemented using natural language processing techniques. Another challenge faced is that sometimes hate may be disguised as humour or even sarcasm. Hence it is important to understand the contextual meaning.

In a study conducted between people who use social media actively and those who do not, it was seen that social media users are subject to hate speech at higher levels as compared to non-users. It was observed that social media interactions bear increased hate speech interactions as compared to direct interactions [2]. Hate speech is also sometimes referred to as cyber-bullying. Cyberbullying has been explored even in languages like Arabic [3][4]. Later it evolved to the term hate speech [5][6]. Hate speech is an umbrella term used to refer to abuse, profanity, sexism.

A. Definition

Various definitions have been given to hate speech which has evolved and been modified over the years. The content that has been referred to as 'hateful' varies according to the researchers and the datasets they are using. It

Grenze ID: 01.GIJET.7.2.13

may also be referred to as a general term for various types of offending user related content [7].
Fortuna & Nunes: “Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used” [8].
Nockleby: “Any communication that disparages a person or a group on the basis of some characteristics such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic” [9].
Facebook: “Direct attack against people on the basis of protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease” [10].
Twitter: “Hateful conduct- You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories” [11].
Snapchat: “Hate speech or content that demeans, defames, or promotes discrimination or violence on the basis of race, colour, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight, or pregnancy status is prohibited” [12].
YouTube: “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” [13].

II. Related Work

Many works have been done on methods that automatically detect hate speech. During the process of literature review several survey papers were found that summarises these works. Major areas that have been investigated to automatically recognize hate speech using natural language processing have been described [7]. They discuss the terminology for the types of offending user-created content which are referred to in the individual works they contain. Features that distinguish the different approaches are analysed. Then they focus on research on individuals involved in hate speech situations and their influence. Some of the works on anticipating alarming societal changes and their significance are discussed. Classification methods, data and annotation for experiments are explored. Finally, shortcomings in existing works which can be considered for future research are pointed out.
Another survey that elaborates on the recent progress in the field, by providing an insight regarding previous approaches, algorithmic methods, and main features used is [8]. They highlight the factors that motivate conducting research in this area. They explore the conceptual aspects of reviewing the topic, differentiate among the various definitions of hate speech, examine certain cases of the concept, associate hate speech with other related concepts, determine evolution of online hate speech, and who are mainly targeted by it. They give a summary of features explored in previous works. They set forth similar open-source projects and datasets. Insights into main challenges and possible research works are also provided.

III. Tools Used

The NLTK library of python is most commonly used to implement natural language processing techniques like stemming, tokenization, parsing etc. TextBlob is an extension of NLTK and allows us to access its facilities in an easier manner. It is more suitable for smaller projects. Another one is SpaCy which is an open-source software library. It is faster, smoother, and more efficient. It also supports API and support for more than 53 languages which is less than NLTK. Another library of python used is gensim whose design goal is memory efficiency and can process arbitrarily large corpora. Standard CoreNLP supports advanced NLP techniques like parts of speech tagger (POS tagger), named entity recognizer (NER), parser, bootstrapped pattern learning, quote attributions and so on. It also supports deep learning-based sentiment analysis tools. But the disadvantage is that CoreNLP is computationally more expensive than other libraries.
OpenNLP also supports more advanced text pre-processing services. It is developed by volunteers and will naturally face some problems like erratic update schedules. Some tool kits are language specific like Language Technology Platform (LTP) which is an open-source Chinese natural language processing tool which uses XML to transfer data through modules. It supports techniques like semantic role labelling, semantic dependency parsing etc. AllenNLP is a deep learning library which is mainly used for research purposes as it supports

scalable development, helps in the management of experiments, and provides the facility of not having to build from scratch. fastText is a library which facilitates learning of word embeddings by using neural networks. It helps in word representations and sentence classifications.

IV. Features

TF-IDF is a characteristic feature used for text classification. Among logistic regression, balanced SVM and GBDT, the word TF-IDF method outperforms character n-gram method [1]. Various efforts have been undertaken to understand the effects of various linguistic features in improving the accuracy. Character n-gram outperforms word n-grams, as character n-gram matrices are sparse compared to word n-gram matrices. Word-length distributions do not contribute much to improving performance, and seldom improve over character-level features [14]. Comparison of features like word n-gram with values of n as 1 and 2, character n-gram with values of n as 3 and 4, and negative sentiment has also been analysed [15].
Adding enhanced features such as count of misspelled words, count of all words, count of emoji, count of special punctuations, the percentage of capitalisation and hashtags, leads to incremental improvement over SVM with basic features like surface features, linguistic features, and sentimental features [16]. Other feature that can be added include URLs. Information regarding a user can also be used to detect hate speech. These include user's tendency to post hateful tweets [6], geographic location of user, his activity engagement on social media sites, personal information, preferences etc. Further when background information of users was added the accuracy further increased.

V. Classifiers

A. Machine Learning

Machine learning models are most commonly employed to detect hate speech. For machine learning methods, supervised learning methods were the most commonly used method. Classifiers like SVM, Logistic regression are most commonly used for classification, while classifiers like Random Forest may be used for categorization [17]. fastText also outperformed word2vec and doc2vec algorithms on 10000 data samples [18].

B. Deep Learning

Deep Learning Methods outperform other baseline methods. The accuracy further can be increased by combining them with gradient boosted random trees [1]. Different configurations of CNN models used have shown better performance as compared to Logistic Regression [19]. When word embeddings were initialised with either random embeddings or GloVe embeddings, CNN performed better than LSTM which was better than fastText. Best results were obtained when tweet embeddings were initialized to random vectors, LSTM was trained using backpropagation. Learned embeddings were then used to train a GBDT classifier. CNN is an efficient network which performs as a 'feature extractor', while orderly sequence learning program can be modelled using RNN [20].
Compared against SVM model, CNN+GRU network and CNN network, the CNN+GRU model with drop-out, global max pooling layers and elastic net regularisation in [16] achieves the highest F1 on all datasets. Inaccurate classification is done by classifiers on hate tweets as compared to non-hate. When a one-step classification was attempted, the HybridCNN model proposed in [21] outperformed WordCNN which was better than CharCNN. But when two logistic regression classifiers were combined for a two-step classification, it had performance equal to that of HybridCNN. fastText classifier exceeds deep learning techniques in terms of speed and accuracy. fastText was compared with other text classifiers for sentiment analysis. Further its evaluation on the tag prediction dataset was done to check its ability to scale [22].

C. Ensemble Classifiers

Some classifiers use a combination of classifiers to do classification. This is based on the fact that some classifiers perform well in certain datasets while others perform well on some others. 5 different classifiers were trained [23] which included- Support Vector Machines, Naïve Bayes, Random Forest, K-Nearest Neighbours and Maximum Entropy, and further some of these classifiers were combined using 2 methods: -
(i)Hard voting-From each classifier a vote is taken. The final selected category would be the one with the majority i.e., greater than half of the votes.
(ii)Soft voting- The final selected category would be the one which has the highest voting score or average probability from each classifier.

VI. People Targeted

Hate speech against women, adolescents, particular individuals, Muslims, and other minority communities are identified. When individuals are targeted, they will usually be tagged by their username. The rest of abuse usually comes under generalised abuse, and they are identified by the vocabulary in the text. The hate speech may also be implicit or explicit. Implicit abuses take on the form of sarcasm, humour which are more difficult to identify [24]. Sexism exhibited online may also be implicit or explicit. The less pronounced forms of sexism exhibited online were also investigated, by classifying tweets as hostile (tweet exhibits explicitly negative emotion and is sexist), benevolent (tweet indicates positive sentiment but is sexist) and others [25].

Some hate tweets are easier to identify compared to the rest. Racist and homophobic tweets mostly tend to be categorized as hate speech, but sexist tweets are categorized as offensive. Tweets lacking explicit hate keywords pose greater difficulty to be classified. Automated classification methods tend to achieve higher accuracy at differentiating between different classes. While analysing the results it was observed that the inclusion or exclusion of certain offensive terms can either increase or decrease classification accuracy. While the presence of certain terms may result in the hateful text correctly being classified, the absence of those terms may result in the hateful text not being identified as hateful. That is a case of misclassification [26]. Without the presence of some keywords, it is difficult to identify the groups being targeted [5].

VII. Dataset

Criteria existing in critical race theory, have been provided in [14] which have been used to annotate a corpus of about 16k tweets. Their dataset consists of 136,052 tweets of which 16,914 tweets were annotated. 3,383 tweets were labelled as sexist, 1,972 as racist, and 11,559 as neither sexist nor racist. One of the largest datasets obtained by crowdsourcing is available in [27] where tweets were labelled with abuse related labels. Dataset focusing on religion (Muslim) and refugees was published by [16]. Hate tweets in the Indonesian Language from the beginning of August 2017 to September 2017 were gathered in [28]. It included a total of 4,002 tweets. A corpus of 17,567 comments were built, which were retrieved using Graph API from 99 Facebook public pages which included Italian newspapers, politicians, artists, and groups [29]. Further details regarding the different datasets available, are given in table I.

VIII. Languages

Tweets also tend to be posted in languages other than English and hence obviously some of those tweets also tend to be hateful. Hate speech detection in some languages tends to be more explored than the others.

One such highly explored language is the Indonesian Language. Indonesia is ranked third in terms of twitter users. Indonesia is diverse in terms of its race, religion, and ethnicity. Hence it consists of a large number of hateful tweets. One such work focused on the dataset, where the dataset consisted of hate speech tweets in general, i.e., targeting religion, ethnicity, race, and gender. It focused on tweets related to the Jakarta Governor Election 2017. The election triggered hate tweets as one of its candidates was a person from the minority community, hence religion and race hate speech tweets were targeted towards that candidate, while the other was a woman who was the target of gender hate speech tweets [7].

Less frequently used languages tend to have less tools available than others and it is difficult to adapt the common language tools to those languages. The challenge associated is that for ethnic groups with scarce online computational resources, it is difficult to gather and process online texts, in addition to detecting hate speech associated with them. This was explored in [32] where Amharic language in Ethiopia was used. Using domain specific embeddings on English Hindi code mixed tweets improved results [30]. Work has also been done on Italian and German tweets where comparative evaluation was proposed on English, Italian and German tweets by using the Neural network that performs well across the three of them [33].

Some languages tend to be more complex than others. Arabic language is one of the languages which is difficult to understand and draw meanings from. This is mainly because of the variations in dialect, meaning across different regions. Variations in handwriting are also present like writing from right to left, combining with other languages like Latin and so on [34]. Sometimes tweets tend to be posted in bi-lingual or multi-lingual form like in [18] which worked on detecting hate speech in Hindi-English code-mixed text.

TABLE I. DATASETS

Dataset	People Targeted	No of Tweets	Labels	Language	Source
Alfina et al. [7]	Minority group (race, religion) Women	1100	HS Non_HS	Indonesia	Twitter
Kamble et al. [30]	Religious Minority, Caste Minority, Sexual Minority	255309	Hate Non-Hate	Hindi-English	Twitter
Jha et al. [25]	Sexist tweets towards men and women	10095	Benevolent Hostile Others	English	Twitter
Gamback et al. [19]	-	6655	Racism Sexism Both (Racism and Sexism) Non-Hate Speech	English	Twitter
Waseem&Hovy[14]	Religious Minority, Ethnic Minority, Sexual Minority	16914	Racism Sexism Neither	English	Twitter
HASOC [31]	No specific targets	7551	Hate and Offensive (HOF) Non- Hate and offensive (NOT)	English, German, Hindi	Twitter, Facebook

IX. CONCLUSION

The paper discussed the need for hate speech detection and the various variations in definition referred to by "hate speech". Further various tools available for hate speech detection were discussed and the scenarios under which each is suitable to be used along with the limitations of each. The features which have been included in the study of detection of hate speech include common features like bag of words, term frequency inverse document frequency, character n-grams as well as more complex features like analysing user behaviour and deriving additional features from tweets. Performance comparison of different classifiers was also analysed.
Common categories of people targeted were also identified from the literature reviews done. It was also understood that in certain scenarios it becomes difficult to identify hate speech if certain keywords are absent. From the literature review it was also understood that English is among the highly researched languages for detecting hate speech. Research in other languages is difficult due to current lack of research, difficulty in collecting tweets as they tend to be scarce. Another challenge of targeting other languages is because some languages are complex, and rigid knowledge is required. In spite of this, work has been published in languages like Italian, German, Arabic etc. Bi-Lingual and multilingual tweets have also been considered in certain cases.

REFERENCES

[1] Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vasudeva Varma. "Deep learning for hate speech detection in tweets." In Proceedings of the 26th international conference on World Wide Web companion, pp. 759-760. 2017.
[2] Barnidge, Matthew, Bumsoo Kim, Lindsey A. Sherrill, Žiga Luknar, and Jiehua Zhang. "Perceived exposure to and avoidance of hate speech in various communication settings." Telematics and Informatics 44 (2019): 101263.
[3] Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu. "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, pp. 71-80. IEEE, 2012.
[4] Haidar, Batoul, Maroun Chamoun, and Fadi Yamout. "Cyberbullying detection: A survey on multilingual techniques." In 2016 European Modelling Symposium (EMS), pp. 165-171. IEEE, 2016.
[5] Warner, William, and Julia Hirschberg. "Detecting hate speech on the world wide web." In Proceedings of the second workshop on language in social media, pp. 19-26. 2012.
[6] Pitsilis, Georgios K., Heri Ramampiaro, and Helge Langseth. "Effective hate-speech detection in Twitter data using recurrent neural networks." Applied Intelligence 48, no. 12 (2018): 4730-4742.
[7] Schmidt, Anna, and Michael Wiegand. "A survey on hate speech detection using natural language processing." In Proceedings of the fifth international workshop on natural language processing for social media, pp. 1-10. 2017.
[8] Fortuna, Paula, and Sérgio Nunes. "A survey on automatic detection of hate speech in text." ACM Computing Surveys (CSUR) 51, no. 4 (2018): 1-30.
[9] Warner, William, and Julia Hirschberg. "Detecting hate speech on the world wide web." In Proceedings of the second workshop on language in social media, pp. 19-26. 2012.

[10] Facebook. Retrieved from https://www.facebook.com/communitystandards/objectionable_content.
[11] Twitter. Retrieved from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.
[12] Snapchat. Retrieved from https://snap.com/en-US/community-guidelines.
[13] Youtube. Retrieved from https://support.google.com/youtube/answer/2801939?hl=e.n&ref_topic=9282436.
[14] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In Proceedings of the NAACL student research workshop, pp. 88-93. 2016.
[15] Alfina, Ika, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. "Hate speech detection in the Indonesian language: A dataset and preliminary study." In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 233-238. IEEE, 2017.
[16] Zhang, Ziqi, David Robinson, and Jonathan Tepper. "Detecting hate speech on twitter using a convolution-gru based deep neural network." In European semantic web conference, pp. 745-760. Springer, Cham, 2018.
[17] Vega, Luis Enrique Argota, Jorge Carlos Reyes-Magaña, Helena Gómez-Adorno, and Gemma Bel-Enguix. "MineriaUNAM at SemEval-2019 Task 5: Detecting hate speech in Twitter using multiple features in a combinatorial framework." In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 447-452. 2019.
[18] Sreelakshmi, K., B. Premjith, and K. P. Soman. "Detection of Hate Speech Text in Hindi-English Code-mixed Data." Procedia Computer Science 171 (2020): 737-744.
[19] Gambäck, Björn, and Utpal Kumar Sikdar. "Using convolutional neural networks to classify hate-speech." In Proceedings of the first workshop on abusive language online, pp. 85-90. 2017.
[20] Ordóñez, Francisco Javier, and Daniel Roggen. "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition." Sensors 16, no. 1 (2016): 115.
[21] Park, Ji Ho, and Pascale Fung. "One-step and two-step classification for abusive language detection on twitter." arXiv preprint arXiv:1706.01206 (2017).
[22] Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. "Bag of tricks for efficient text classification." arXiv preprint arXiv:1607.01759 (2016).
[23] Fauzi, M. Ali, and Anny Yuniarti. "Ensemble method for indonesian twitter hate speech detection." Indonesian Journal of Electrical Engineering and Computer Science 11, no. 1 (2018): 294-299.
[24] Waseem, Zeerak, Thomas Davidson, Dana Warmsley, and Ingmar Weber. "Understanding abuse: A typology of abusive language detection subtasks." arXiv preprint arXiv:1705.09899 (2017).
[25] Jha, Akshita, and Radhika Mamidi. "When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data." In Proceedings of the second workshop on NLP and computational social science, pp. 7-16. 2017.
[26] Davidson, Thomas, 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, vol. 11, no. 1. 2017.
[27] Founta, Antigoni, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. "Large scale crowdsourcing and characterization of twitter abusive behavior." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, no. 1. 2018.
[28] Putri, T. T. A., S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean. "A comparison of classification algorithms for hate speech detection." In IOP Conference Series: Materials Science and Engineering, vol. 830, no. 3, p. 032006. IOP Publishing, 2020.
[29] Del Vigna12, Fabio, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. "Hate me, hate me not: Hate speech detection on facebook." In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), pp. 86-95. 2017.
[30] Kamble, Satyajit, and Aditya Joshi. "Hate speech detection from code-mixed hindi-english tweets using deep learning models." arXiv preprint arXiv:1811.05145 (2018).
[31] Mandl, Thomas, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. "Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages." In Proceedings of the 11th forum for information retrieval evaluation, pp. 14-17. 2019.
[32] Mossie, Zewdie, and Jenq-Haur Wang. "Vulnerable community identification using hate speech detection on social media." Information Processing & Management 57, no. 3 (2020): 102087.
[33] Corazza, Michele, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. "A multilingual evaluation for online hate speech detection." ACM Transactions on Internet Technology (TOIT) 20, no. 2 (2020): 1-22.
[34] Al-Hassan, Areej, and Hmood Al-Dossari. "Detection of hate speech in social networks: a survey on multilingual corpus." In 6th International Conference on Computer Science and Information Technology, vol. 10. 2019.