iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41245

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Un-Compromised Credibility: Social Media based Multi-Class Hate Speech Classification for Text: A Review

Miss. Priyanka R Telshinge[1], Mr. Mangesh D Salunke[2]
[1]*PG Student, Department of Computer Engineering, Rajarshee Shahu Institute of Technology and Research, Savitribai Phule Pune University, Pune - 411041, India*
[2]*Assistant Professor, Department of Computer Engineering, Rajarshee Shahu Institute of Technology and Research, Savitribai Phule Pune University, Pune - 411041, India*

Abstract: ***Hate speech is a crime that has been on the rise in recent years, not just in face-to-face contacts but also online. This is due to a number of causes. On the one hand, due of the anonymity given by the internet and social networks in particular, people are more likely to engage in hostile behaviour. People's desire to voice their thoughts online, on the other side, have increased, adding to the spread of hate speech. Governments and social media platforms can benefit from detection and prevention techniques because this type of prejudiced speech can beimmensely destructive to society. We contribute to a solution to this dilemma by giving a systematic review of research undertakenin the subject through this survey. This challenge benefited from the use of several complicated and non-linear models, and CAT Boost performed best due to the application of latent semantic analysis (LSA) for dimensionality reduction.***

Keywords: Multi-Class Hate Speech, Natural Language Processing, Hate Speech Classification, Social Media Micro blogs, Multi-Class Hate SpeechDataset.

I. INTRODUCTION

Online social network (OSN) is the use of dedicated websites applications that allow users to interact with other users or to find people with similar own interest Social networks sites allow people around the world to keep in touch with each other regardless of age [1] [7]. Sometimes children are introduced to a bad world of worst experiences and harassment. Users of social network sites may not be aware of numerous vulnerable attacks hosted by attackers on these sites. Today the Internet has become part of the people daily life. People use social networks to share images, music, videos, etc., social networks allows the user to connect to several other pages in the web, including some useful sites like education, marketing, online shopping, business, e-commerce and Social networks like Facebook, LinkedIn, Myspace, Twitter are more popular lately [8][9]. The offensive language detection is a processing activity of natural language that deals with find out if there are shaming (e.g. related to religion, racism, defecation, etc.) present in a given document and classify the file document accordingly [1]. The document that will be classified in abusive word detection is in English text format that can be extracted from tweets, comments on social networks, movie reviews, and political reviews. Hate speech is a crime that has been on the rise in recent years, not just in face-to-face contacts but also online. This is due to a number of causes. On the one hand, due of the anonymity given by the internet and social networks in particular, people are more prone to engage in hostile behaviour, People, on the other hand, are more willing to share their thoughts online, which contributes to the spread of hate speech as well. Governments and social media platforms can benefit from detection and prevention techniques because this type of prejudiced speech can be immensely destructive to society. We contribute to a solution to this dilemma by giving a thorough overview of research undertaken in this area through this survey.

Hate speech is defined as a discourse that is potentially hurtful to a person's or group's feelings and may contribute to violence or insensitivity, as well as irrational and inhuman behaviour. Hate speech has increased as a result of the growth of online social media, which is illegal. Hate speech and hate crimes are linked, and there is evidence that hate crimes are on the rise. As the problem of hate speech grows in popularity, many government-led initiatives are being implemented, such as the Council of Europe's No Hate Speech movement. Legislation has also been enacted to combat its spread, dubbed the EU Hate Speech Code of Conduct, which must be signed and implemented by all social media services within 24 hours.

The majority of the issues raised are primarily connected to the dataset's quality, which will be addressed in this study through the creation of quality-based strong datasets. The second problem, which is also addressed in this paper, is to investigate and determine the bestset of characteristics for hate speech identification before developing a suitable classifier. When looking at the FBI's hate crime data, the most common categories are based on race, ethnicity, and religion. As a result, all of these categories are largely chosen for the production of datasets.

II. LITERATURE REVIEW

Sr No	Paper Details	Advantages	Algorithm/ Techniques	Limitations	Summary
1	Fortuna, Paula, and Sérgio Nunes. "A survey on automatic detection of hate speech in text." ACM Computing Surveys (CSUR) 51.4 (2018): 1-30	The development and systematization of shared resources, such asguidelines, annotated datasets in multiple languages, and algorithms, is a crucial step in advancing the automatic detection of hate speech.	CCS Concepts: Natural language processing; Information extraction; Information systems;Sentiment analysis Algorithm: Hate Speech Detection	There are not many studies and papers published in automatic hate speech detection from a computer scienceand informatics perspective. This slows down the progress of the research, because less datais available, making it more difficult to compare results from different studies	In this paper, we provided a critical assessment of how automatic detection of hate speech in text has grown over the years in this survey. First, we looked at hate speech in many circumstances, rangingfrom social media platforms to other organisations.
2	Kumar R, Ojha AK, Malmasi S, Zampieri M. Benchmarking Aggression Identification in Social Media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). ACL; 2018. p. 1–11.	The performance of the best systems in the task shows that aggression identification is a hard problem to solve.	Aggression Identification organized with theTRAC workshop atCOLING 2018	We find quite a few neural networks-based systems not performing quite well in the task	In this paper, we have presented the report of the First Sharedtask on Aggression Identification organized with the TRAC workshop at COLING 2018. The shared task received a very encouraging response from the community which underlines the relevance and need of the task. More than 100 teams registered and 30 teams finally submitted their system.
3	de Gibert O, Perez N, Garc'ia-Pablos A, Cuadros M. Hate Speech Dataset from a White Supremacy Forum. In: 2nd Workshop on Abusive	This paper provides thoughtful qualitative and quantitative study of the resulting dataset and several baselin e	Hate speech dataset obtained fro mStormfront	A custom annotation tool has been developed to carry out the manual labelling task which, among other things,	This research provides a hate speech dataset that was manually labelled and collected from Stormfront, a white supremacist online community.
	Language Online@EMNLP; 2018	experiments with different classification models. The dataset is publicly available		allows the annotators to choose whether to read the context of a sentence before labelling it	
4	Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." Proceedings of the International AAA IConference on Web and Social Media. Vol. 11. No. 1. 2017.	This method can achieve relatively high accuracy.	crowd-sourced hat espeech	Tweets without explicit hate keywords are also more difficult to classify.	If we conflate hate speech and offensive language then we erroneously consider many people to be hate speakers and fail differentiate between commonplace offensiv e language and serious hate speech.

5	Unsvåg, Elise Fehn, and Björn Gambäck. "The effects of user features on twitter hate speech detection." Proceedings of the 2nd workshop on abusive language online (ALW2). 2018	It improves the baseline classifier performance.	Logistic Regression-based hate speech, N--gram	They were developed for different subtasks and languages, with different geographical areas of the users in the datasets, and in particular with different interpretations an dannotations of hate speech.	, The article focused on Twitter to In this paper investigate the possibility and implications of adding user attributes in hate speech classification.
6	Vu, Xuan-Son, et al. "HSD shared task in VLSP campaign 2019: Hate speech detection for social good." arXiv preprint arXiv: 2007.06493 (2020). .	The social network data to better support society in the information age for the next VLSP campaign in 2020.	Hate Speech Detection (HSD)	Security is less	In this paper, The Hate Speech Detection (HSD) shared task in the VLSP Campaign 2019 has been a valuable exercise in building predictive models to filter out hate speech contents on social networks.
7	Mathew, Binny, et al. "Analyzing the hate and counter speech accounts on twitter." arXiv preprint arXiv:1812.02712 (2018)	• It is faster. • It is more flexible and responsive. • It is capable of dealing with extremism from anywhere and in any language. • It does not form a barrier against the principle of free and open public space for debate.	Supervised model	No efficient	In this paper, we perform the first characteristic study comparing the hateful and counter speech accounts in Twitter. We provide a dataset of 1290 tweet-reply pairs of hate speech and the corresponding counter speech tweets.
8	Gaydhani, Aditya, et al. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach." arXiv preprint arXiv: 1809.08651 (2018).	In this Logistic Regression performs better with the optimal n-gram range 1 to 3 for the L2 normalization of TFIDF. Accuracy is more.	Logistic Regression, Naive Bayes and Support Vector Machines algorithms, TFIDF	It was seen that the model does not account for negative words present in a sentence.	In this paper, we proposed a solution to the detection of hate speech and offensive language on Twitter through machine learning using n-gram features weighted with TFIDF values.
9	Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE access 6 (2018): 13825-13835.	Automatically detects hate speech patterns	hate speech patterns and most common unigrams and use these along with sentimental an d semantic features to classify tweets into hateful, offensive and clean.	The accuracy is less.	In this paper, In order to detect hate on Twitter, we proposed a novel approach. Our proposed method classifies tweets into hateful, offensive, and clean categories by automatically detecting hate speech patterns and the most common unigrams, as well as emotive and semantic aspects
10	Wich, Maximilian, Jan Bauer, and Georg Groh. "Impact of politically biased data on hate speech classification." Proceedings of the Fourth Workshop on Online Abuse and Harms. 2020.	To identify bias with XAI in existing data sets or during data collection. To use these findings to build politically branded hate speech filters that are marked as those.	ML models, unbiased data sets.	we simulate the politicalbias and construct synthetic data sets with offensive tweets annotated by humans and non-offensive tweets that are only implicitly labelled. The GermEval data andour gathered data are from different periods	, we found an indication that the degree of impairment might depend on the political orientation of bias. we provide a proof-of-concept of visualizing such a bias with explainable ML models. The results can help to build unbiased data sets or to debias them

III. OPEN ISSUES

Lot of work has been done in this field because of its extensive usage and applications. In this section, some of the approaches which have been implemented to achieve the same purpose are mentioned. These works are majorly differentiated by the techniques for multi- keyword search and group sharing systems.

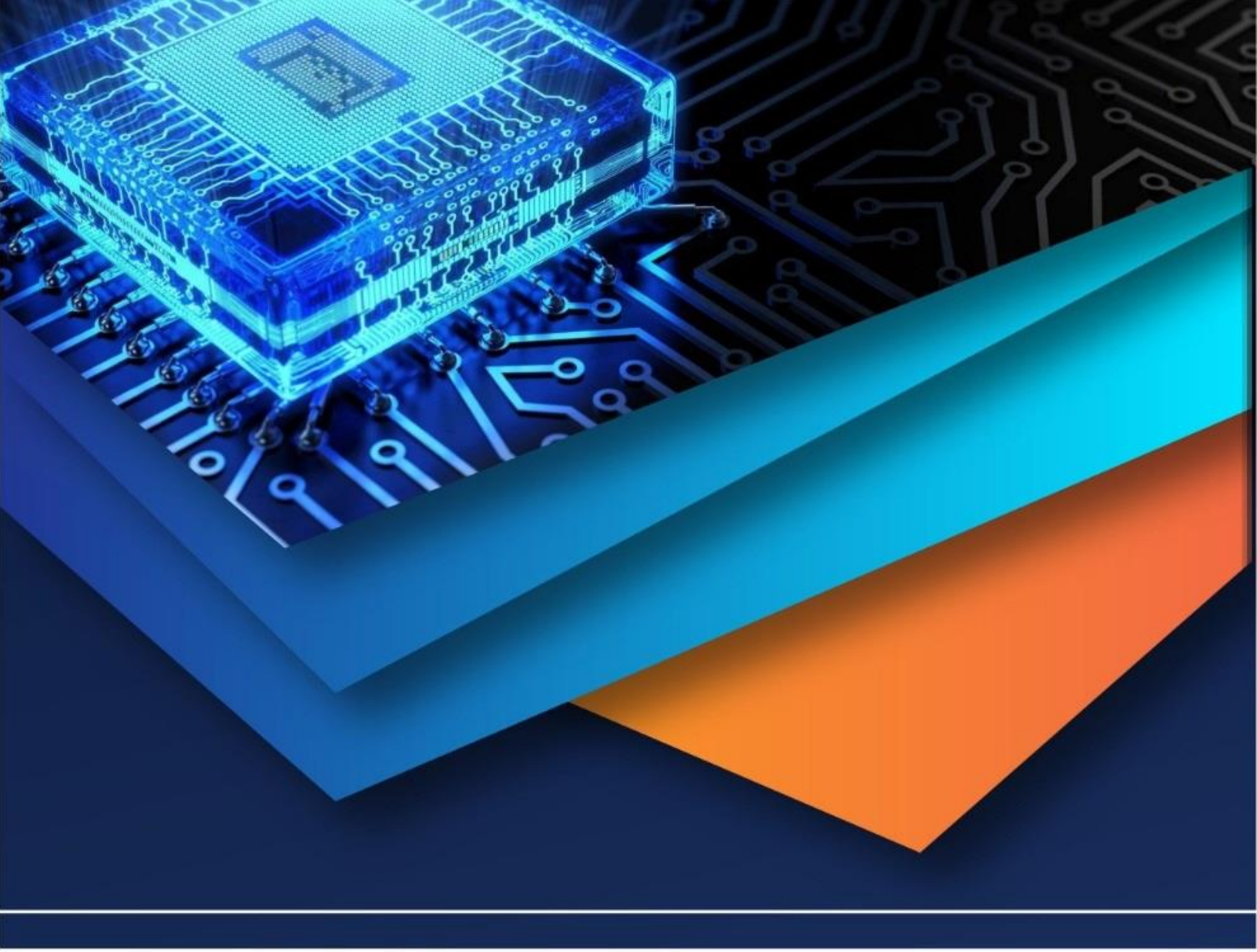
1) In previous technology in which A Survey on Automatic Detection of Hate Speech in Tex word sequence was ignored.
2) In White Supremacy Forum, The dataset is unbalanced as there exist many more sentences not conveying hate speech than 'hateful" ones.
3) The Effects of User Features on Twitter Hate Speech Detection, this subset improvement may have been affected by the individual feature(number of) 'Followers', which also increased the F1-score on the two datasets.
4) The proposed sets of unigrams and patterns can be used as already-built dictionaries not included it is used for future works related to hate speechdetection.

IV. CONCLUSION

The complex problem of multi-class automated hate speech categorization for text is solved with considerably better results after the primary challenges are discovered first. There are ten unique binary categorised datasets made up of various hate speech categories. Experts annotated each dataset with a high level of agreement among annotators, using a set of detailed, well-defined guidelines. The datasets were well-balanced and comprehensive. They were also enriched with linguistic nuance. Compilation of such a dataset was accomplished as an essential need for filling the field's gap.

REFERENCES

[1] Fortuna, Paula, and Sérgio Nunes. "A survey on automatic detection of hate speech in text." ACM Computing Surveys(CSUR) 51.4 (2018): 1-30.
[2] Kumar R, Ojha AK, Malmasi S, Zampieri M. Benchmarking Aggression Identification in Social Media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). ACL; 2018. p. 1–11.
[3] de Gibert O, Perez N, Garc'ia-Pablos A, Cuadros M. Hate Speech Dataset from a White Supremacy Forum. In: 2nd Workshop on Abusive Language Online@EMNLP; 2018.
[4] Davidson, Thomas, et al. "Automated hate speech detectionand the problem of offensive language." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017.
[5] Unsvåg, Elise Fehn, and Björn Gambäck. "The effects of user features on twitter hate speech detection." Proceedings of the 2nd workshop on abusive language online (ALW2).2018.
[6] Vu, Xuan-Son, et al. "HSD shared task in VLSP campaign 2019: Hate speech detection for social good." arXiv preprint arXiv:2007.06493 (2020).
[7] Mathew, Binny, et al. "Analyzing the hate and counterspeech accounts on twitter." arXiv preprint arXiv:1812.02712 (2018).
[8] Gaydhani, Aditya, et al. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach." arXiv preprint arXiv: 1809.08651 (2018).
[9] Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE access 6 (2018): 13825-13835.
[10] Wich, Maximilian, Jan Bauer, and Georg Groh. "Impact of politically biased data on hate speech classification." Proceedings of the Fourth Workshop on Online Abuse and Harms. 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY