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

Miss. Priyanka R Telshinge[1], Dr. Mangesh D Salunke[2]
[1]*PG Student, Department of Computer Engineering, JSPM Narhe Technical Campus,Pune, India*
[2]*Assistant Professor, Department of Computer Engineering, JSPM Narhe Technical Campus,Pune, India*

***Abstract*— Not just in person but also via the internet, people are committing more crimes related to hate speech, which has been on the rise in recent years. A variety of factors have contributed to this result. People are more inclined to participate in hostile behaviour online because of the anonymity provided by the internet and social networks in particular. On the other hand, people are more likely to engage in hostile behaviour offline. On the other hand, more and more people are using the internet to share their opinions, which contributes to the proliferation of hate speech. Because this kind of prejudiced speech has the potential to be so detrimental to society, implementing detection and prevention strategies can be beneficial for governments as well as social media companies. By providing a comprehensive evaluation of the research that has been carried out on the topic through the course of this survey, we make a contribution toward resolving this conundrum. This task benefited from the employment of multiple sophisticated and non-linear models, and CAT Boost fared the best as a result of the application of latent semantic analysis (LSA) for the purpose of dimensionality reduction.**

Index Terms: **Hate Speech, Natural Language Processing, Classification, Social Media Micro blogs, Twitter Dataset.**

I.INTRODUCTION

The term ”online social network” (OSN) refers to the use of specialised websites and programmes that enable users to communicate with one another or to locate other individuals who share their interests. Social networking websites make it possible for people of all ages to stay in touch with one another no matter where they are in the world [1][7]. There are times when youngsters are exposed to the harshest experiences and harassment that the world has to offer. It’s possible that users of social networking sites are unaware of the countless attacks that are being hosted by cybercriminals on these sites as vulnerable vectors. The Internet has quickly integrated itself into people’s everyday lives today. People use social networks to share media such as images, music, videos, and other types of content. Social networks also enable users to connect to a variety of other pages on the internet, including some helpful websites such as those for education, marketing, online shop- ping, business, and e-commerce. These days, social networking sites such as Facebook, LinkedIn, Myspace, and Twitter are becoming increasingly common [8][9]. The detection of offensive language is a natural language processing activity that seeks to determine whether or not a given document contains shaming language (such as language related to religion, racism, or defecation, for example) and then classifies the file document in accordance with this determination [1]. The document that will be categorised in abusive word detection is in the English text format. This document can be taken from tweets, comments on social networks, movie reviews, and political reviews.

Online and in person, hate speech has become a felony that is on the rise in recent years. Several factors are to blame. People are more likely to engage in hostile behaviour because of the anonymity provided by the internet and social networks in particular, but they are also more inclined to communicate their ideas online, which helps to the spread of hate speech. Anti-prejudice measures benefit governments and social media platforms alike, because hate speech like this has the potential to do enormous harm to society. Surveying the field of research in this area helps us find a solution to the problem at hand.

A discourse that has the potential to be painful to the feelings of a person or group and that may contribute to acts of violence or insensitivity, as well as

behaviours that are unreasonable and inhuman, is considered to be hate speech. The proliferation of illegally accessed social media platforms online has led to a surge in the dissemination of hate speech. There is a correlation between hate speech and hate crimes, and there is mounting evidence to suggest that hate crimes are on the rise. As awareness of the problem of hate speech increases, numerous government-driven initiatives are being put into action. One such initiative is the No Hate Speech movement, which is being spearheaded by the Council of Europe. In addition, legislation has been enacted to counteract its spread. This legislation is known as the EU Hate Speech Code of Conduct, and it stipulates that all social media platforms must sign and apply it within twenty-four hours. This study will address the majority of the problems that have been brought up, the majority of which are primarily tied to the quality of the dataset. This problem will be addressed by the establishment of quality-based strong datasets. The second difficulty, which is also addressed in this work, is to research and decide the best collection of features for identifying hate speech before building an appropriate classifier. This problem is investigated and determined in this paper. When looking at the data on hate crimes collected by the FBI, the most common categories are those that are based on religion, race, and ethnicity. As a direct consequence of this, each of these categories is selected for use in the production of datasets to a significant degree.

II.REVIEW OF LITERATURE

Fortuna, Paula, and Se´rgio Nunes. ”A survey on automatic detection of hate speech in text.” ACM Computing Surveys (CSUR) 51.4 (2018): 1-30: In this survey, we gave a critical evaluation of how automatic detection of hate speech in text has developed over the years through this study. The survey was carried out over a period of several years. First, we investigated the concept of hate speech in a variety of contexts, spanning from social media platforms to other types of organisations.

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.: At COLING 2018, we presented the report of the First Shared task on Aggression Identification, which was organised in conjunction with the TRAC workshop. A very promising reaction was obtained from the community in regard to the shared task, which highlights the importance and relevance of the task. More than one hundred teams signed up, however only thirty of them ended up submitting their system.

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: This piece of study presents a dataset of hate speech that was manually labelled and collected from the online group Stormfront, which is associated with white supremacist ideology.

Davidson, Thomas, et al. ”Automated hate speech detection and the problem of offensive language.” Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017: If we equate offensive language with hate speech, then we incorrectly perceive a large number of people to be hate speakers and fail to discriminate between everyday offensive language and serious hate speech.

Unsva˚g, Elise Fehn, and Bjo¨ rn Gamba¨ck. ”The effects of user features on twitter hate speech detection.” Proceedings of the 2nd workshop on abusive language online (ALW2). 2018: The article focused on Twitter to In this paper investigate the possibility and implications of adding user attributes in hate speech classification.

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 Hate Speech Detection (HSD) shared task in the VLSP Campaign 2019 has been a valuable exercise in this paper’s construction of predictive models to filter out contents of hate speech on social networks. This task was part of the VLSP Campaign 2019.

Mathew, Binny, et al.” Within the scope of this paper, we conduct the first study of its kind to compare the characteristics of hate speech and counter speech accounts on Twitter. We present a dataset consisting of 1290 tweet-reply pairs that are

examples of hate speech and the corresponding tweets that are examples of counter speech.

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).: Through the use of machine learning with n-gram features that were weighted with TFIDF values, we proposed a solution in this paper to the problem of detecting offensive language and hate speech on Twitter.

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.:In this piece of research, we proposed a novel approach to the problem of identifying hate speech on Twitter. The approach that we have suggested separates tweets into categories that are either hateful, offensive, or clean by automatically identifying patterns of hate speech, the most common unigrams, as well as emotional and semantic components.

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.: We found evidence to support the hypothesis that the degree of impairment may be influenced by the political persuasion of the bias. We demonstrate the feasibility of visualising such a bias using explainable machine learning models in our proof-of-concept. The findings may be used to construct unbiased data sets or to remove bias from existing ones.

III.PROPOSED METHODOLOGY

In the systemic approach that we have proposed, the task at hand is formulated as a classification of the problem, and its purpose is to detect and mitigate the negative side effects of online public hate speech. The two primary contributions are as follows: 1) The categorization as well as the automatic classification of tweets containing hate speech 2) Construct a web application that will allow users of Twitter to recognise hate speech.

A. Architecture

Fig. 1. Proposed System Architecture

The objective is to have tweets automatically classified into nine different categories. The primary operational components are depicted in figure 1. The preprocessing and feature extraction steps are performed on the labelled training set and test set that correspond to each category.

B. Algorithms

Naive Bayes Steps:

- Given training dataset D which consists of documents belonging to different class say Class A and Class B
- Calculate the prior probability of class A=number of objects of class A/total number of objects
- Calculate the prior probability of class B=number of objects of class B/total number of objects
- Find NI, the total no of frequency of each class Na=the total no of frequency of class A Nb=the total no of frequency of class B
- Find conditional probability of keyword occurrence given a class:
- P (value 1/Class A) =count/ni (A) P (value 1/Class B) =count/ni (B) P (value 2/Class A) =count/ni (A) P (value 2/Class B) =count/ni (B)

  ……

  …….

- P (value n/Class B) =count/ni (B)
- Avoid zero frequency problems by applying uniform distribution
- Classify Document C based on the probability p(C/W)
- a. Find P (A/W) =P (A)*P (value 1/Class A)* P (value 2/Class A) P(value n /Class A)
- b. Find P (B/W) =P (B)*P (value 1/Class B)* P (value 2/Class B). P(value n /Class B)
- Assign document to class that has higher probability.

C. Mathematical Model

The mathematical model for Shamming Detection System is as,

S = {I, F, O }

Where, I = Set of inputs

The input consists of set of Words. It uses Twitter dataset. F

= Set of functions

F = F1, F2, F3. FN

F1: Tweets Extraction F2: Tweets Preprocessing F3: Feature Extraction

F4: Hate Speech Classification O: Hate Speech Detection

IV.RESULTS AND DISCUSSION

Using Twitter application programming interface (API), a large number of tweets are collected. Performance scores for the nine classifiers are shown in Fig 2. The precision and recall scores for the classifiers are given in Table 1.

A. Classification Performance

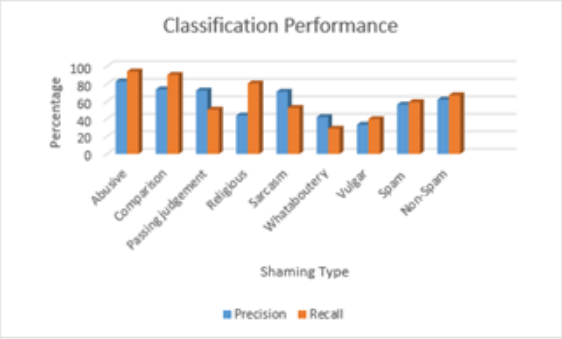

Fig. 2. Classification Performance

Shaming Type	Precision	Recall
Abusive	82.89%	94.2%
Comparison	73.81%	90.4%
Passing judgment	72.42%	50.68%
Religious	44.00%	80.63%
Sarcasm	71.07%	52.67%
Whataboutery	42.12%	28.89%
Vulgar	33.45%	39.83%
Spam	56.43%	59.21%
Non-Spam	62.03%	67.09%

Table 1: Classification Performance

V. CONCLUSION

After the primary challenges have been identified, the difficult problem of multi-class automated hate speech categorisation for text can be solved with significantly improved results. There are ten distinct binary datasets, each of which is categorized according to a different type of hate speech. Each dataset was annotated by specialists, and there was a high level of consensus among the annotators thanks to the use of a comprehensive and clear-cut set of guidelines. The datasets were comprehensive and well-balanced at the same time. In addition to this, linguistic nuance was added to enrich them. In order to meet an essential requirement for closing the gap in knowledge within the field, a dataset of this kind was compiled and compiled successfully.

REFERENCES

[1] KHUBAIB AHMED QURESHI, MUHAMMAD SABIH" Un- Compromised Credibility: Social Media based Multi-Class Hate Speech Classification for Text" IEEE Access 2021

[2] Fortuna, Paula, and Se´rgio Nunes. "A survey on automatic detection of hate speech in text." ACM Computing Surveys (CSUR) 51.4 (2018): 1-30.

[3] 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.

[4] 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.
[5] Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017.
[6] Unsva˚g, Elise Fehn, and Bjo¨ rn Gamba¨ ck. "The effects of user features on twitter hate speech detection." Proceedings of the 2nd workshop on abusive language online (ALW2). 2018.
[7] Vu, Xuan-Son, et al."HSD shared task in VLSP campaign 2019: Hate speech detection for social good." arXiv preprint arXiv:2007.06493 (2020).
[8] Mathew, Binny, et al. "Analyzing the hate and counter speech accounts on twitter." arXiv preprint arXiv:1812.02712 (2018).
[9] 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).
[10] 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.
[11] 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.