Proceedings of the Seventh International Conference on Communication and Electronics Systems (ICCES 2022)
IEEE Xplore Part Number: CFP22AWO-ART; ISBN: 978-1-6654-9634-6

Hate Speech Detection using Machine Learning

P. Preethy Jemima
Assistant Professor, Department of Computer Science and Engineering
SRM Institute of Science and Technology
Ramapuram Campus, Chennai, Tamil Nadu, India
bm9206@gmail.com

Bishop Raj Majumder
Student, Department of Computer Science and Engineering
SRM Institute of Science and Technology
Ramapuram Campus, Chennai, Tamil Nadu, India
bm9206@gmail.com

Bibek Kumar Ghosh
Student, Department of Computer Science and Engineering
SRM Institute of Science and Technology
Ramapuram Campus, Chennai, Tamil Nadu, India
bg1158@srmist.edu.in

Farazul Hoda
Student, Department of Computer Science and Engineering
SRM Institute of Science and Technology
Ramapuram Campus, Chennai, Tamil Nadu, India
fh3170@srmist.edu.in

2022 7th International Conference on Communication and Electronics Systems (ICCES) | 978-1-6654-9634-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCES54183.2022.9835776

***Abstract* – A lot of methods have already been created for the automation of hate speech detection online. There are two elements to this process: identifying the qualities that these terms utilize to target a certain group and classifying textual material as hate or non-hate speech. Due to time restraints, research efforts are initiated on the latter issue in this project. For this reason, detecting hate speech is a more challenging endeavor, as our research of the language used in typical datasets reveals that hate speech lacks distinctive, discriminatory characteristics. Deep neural network topologies are very useful for capturing the meaning of hate speech and are thus proposed as feature extractors. Data from social media sites such as Twitter are used to test the effectiveness of these procedures, and they reveal a 6 percentage point improvement in macro-average F1 or a 9 percent improvement for content that has been labeled as hateful, respectively.**

Keywords— Random Forest, NLP, Machine Learning, Text mining, Sentiment Analysis, Back Propagation Neural Network, ANNs, RNN, LSTN.

I. INTRODUCTION

People may now communicate and share material more easily than ever before, but social media platforms like Twitter and other community forums are also being used to disseminate hate speech and organize events based on hatred [1]. Such media's anonymity and mobility make it easier for hate speech to propagate, which eventually leads to hate crime in a virtual place that traditional law enforcement can't reach. Using the word "hate speech" refers to any kind of communication that disparages someone or a group of people because of their qualities, such as religion, ethnicity, sexual orientation or gender [2]. Recent events, such as Britain's decision to leave the European Union and the terrorist attacks in Manchester and London, have resulted in an uptick in hate speech directed towards immigrant and Muslim populations in the UK. Hate speech is on the rise among young people in the European Economic Area (EEA) area, according to EU polls and studies [3]. According to a survey, 80 percent of respondents had experienced hate speech online, and 40 percent have been intimidated or attacked as a result of their opinions. Since Trump's election, there has been an increase in hate speech and crime in the United States. An increasing number of worldwide efforts have been developed in an effort to better understand the problem and devise effective solutions [4].

II. ORGANISATION

The following is a summary of the content of this document. Section II provides an overview of the topic at issue. System architecture explanations and diagrams are given in Section III. Section IV describes how the system has been implemented, the modules and a few diagrams to explain better the working of the program. And finally Section V concludes the study as well as mentions future enhancements that could be made.

III. PROBLEM STATEMENT

"Tweets" are messages that members post and respond to on Twitter, one of the most popular social networking sites. Individuals can use this to share their ideas and opinions on a variety of topics [5]. The sentiment analysis of such tweets has been used by a wide range of entities, including consumers and marketers, to get insights about products or conduct market study. The accuracy of the sentiment analysis prediction models increases with the recent advances in machine learning techniques [6]. We want to use a variety of machine learning methods to analyze "tweets" in our research. We begin by determining the tweet's polarity, or whether it is good or negative. A tweet with both positive and negative aspects will be classified as either positive or negative depending on which mood is more prominent [7]. According to Kaggle's data set, we are utilizing a crawling dataset that has been classified positive or negative. Emotional expressions and hashtags are also included in the data. As a result, they will need to be processed and then transformed into a format that is standard. We'll also need to sift through the content to find important information [8]. As a type of "tweet," these properties include unigrams and bigrams, which may be used to convey information. Although individual models did not provide a high level of accuracy, we selected a small number of the best models to create an ensemble [9]. To boost the accuracy of prediction, we employ ensembling, a form of meta learning algorithm technique that combines several classifiers. And at the end, we report our findings and experimental results [10].

IV. EXISTING SYSTEM

A document classification task is all that exists in terms of existing methodologies. SVM [12], Naive Bayes [13], and Logistic Regression [14] (traditional approaches) rely on

manual feature engineering; the more contemporary deep learning paradigm, on the other hand, uses deep learning methods or neural networks on raw data to automatically generate multi-layers of abstract features from it [15-17]

A. Disadvantages

1. Most research on hate speech detection have relied on Micro-average Precision, Recall, and F1.

2. The issue here is that micro-averaging might conceal the true performance of minority classes in an unbalanced dataset where instances of one class (to be termed the "dominant class") far outnumber those of other classes (to be called "minority classes")[18-19].

V. PROPOSED SYSTEM

Only 5.8% (DT) to 31.6 percent (WZ) of all tweets are hateful, which means that all datasets are heavily skewed against hatred. 'Racism,' and the extreme situation that carries 'both,' become increasingly uncommon when specific sorts of hate are examined in detail. Aside from that, there are two possible outcomes. Initially, an evaluation measure, such as the micro F1 that examines a system's performance on the full dataset and that is done without consideration to class differences that may be skewed against the system's capacity to identify non-hate, will be used to assess the system's performance. In a nutshell, a theoretical system that achieves virtually flawless F1 in recognizing 'racism' Tweets in identifying 'non-hate' can nevertheless hide a low F1 in identifying 'racism' Tweets. Second, there is a dearth of hate-related data compared to non-hate-related data. Because the numbers were gleaned from Twitter and represent the true nature of data imbalance throughout this sector, it's possible that this is a more difficult topic to discuss than it first appears. Because of this, we will very certainly have to devote a significant amount of time and effort to providing more training data for non-hateful content.

A. Advantages

1. Oversampling or undersampling may not be able to solve this problem, as we'll illustrate in the next section.
2. There are no distinguishing linguistic traits in hate Tweets that are not present in non-hate tweets.

One way of measuring how distinct hate and non-hate Twitter messages are in their vocabulary is to look at how many words are unique to each class.

VI. SYSTEM ARCHITECTURE

A. Data Description:

In the form of comma-separated values files, the data includes tweets and the attitudes they express. Tweet id is a unique identifier, and sentiment is either 1 (positive) or 0 (negative). The tweet is enclosed in "." in a CSV training dataset. CSV files of the form tweet id, tweet are used for the test dataset.

There is a considerable bias towards non-hate in all datasets because hate Tweets account for just 5.8% (DT) to 31.6 % of the total tweets (WZ). Those with "racism" and, as previously indicated, the extreme scenario in which it contains "both" are even fewer when we look at specific sorts of hate. Aside from that, there are two possible outcomes. The system's capacity to detect 'non-hate' may be overstated when evaluated using a metric like the micro F1, which considers a system's performance throughout the whole dataset, independent of class. Racist-tagged tweets can still be obscured by tweets labeled "non-hate," and vice versa, even if a hypothetical system can reach virtually flawless F1 in the recognition of 'racism' tagged tweets. To begin with, there aren't nearly as many hate-related Tweets as there are non-hate-related ones. Due to the fact that the datasets are being obtained from Twitter, this scenario may not be as simple to fix as it appears. As a result, substantially more time and effort will be required to annotate the non-hate component of training data in order to detect as many occurrences of hateful content as possible.

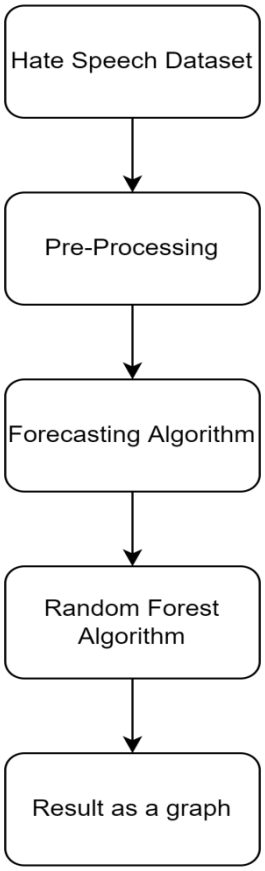

Fig 1: System Design

VII. SYSTEM IMPLEMENTATION

A. Modules:

A.1 Exploratory Data Analysis:

The target variable was established in this stage after some descriptive analysis. Other potentially problematic (high cardinality) factors were also examined, including the target's number of classes. Visualizing the target variable in a histogram, an excellent tool for understanding the distribution of data, was used to aid with parameter tuning.

A.2 Data Cleaning:

As a prelude to pre-processing, the high cardinality variables were removed in this stage.

A.3 Pre-processing & Transformation:

One-hot encoding was used to change the categorical variable from the complete data set and subsequently to make it easier to analyze. When dealing with data in a sparse matrix format, this is a criterion that must be satisfied by specific algorithms. Automated software for statistics such as R can handle this phase when creating models. The data is then populated with zeros for any missing values. Min-max normalization is then used to the continuous variables to translate their values onto a scale from 0 to 1 so that the coefficients are not affected by variables that are on different scales.

A.4 Data Partition:

Partitioning the pre-processed data into training and testing sets is the next step.

A.5 Modelling:

In order to build a k-NN classifier model, 10 neighbor classes and the Euclidean distance between them were used.

A.6 Evaluation:

The classifier is subsequently evaluated using new test data, and the R squared values for the training and test datasets are computed.

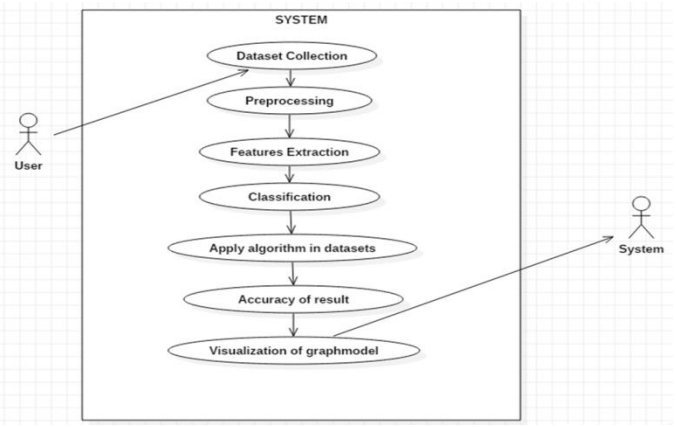

Fig 2: Use Case Diagram

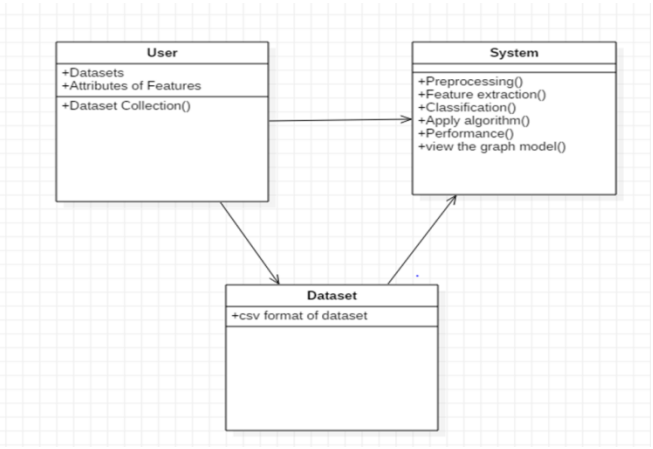

Fig 3: Class Diagram

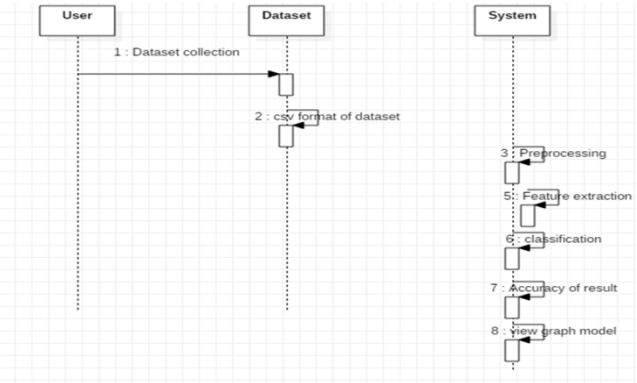

Fig 4: Sequence Diagram

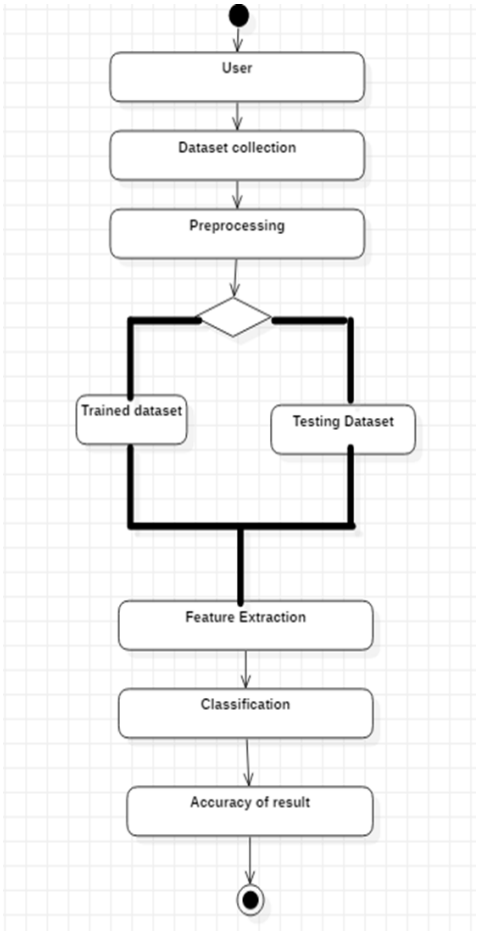

Fig 5: Activity Diagram

VIII. CONCLUSION

We conducted a poll to see if hate speech identification could be automated. The most common way to describe this challenge is as a problem of supervised learning. In a regular order, features that are sufficiently generic, such as word bags or word embeddings, provide good classification performance. Character-level techniques outperform token-level ones. There are several lists of slurs that can aid in categorization, but only when they are used in conjunction with other traits. Many more advanced characteristics, such as rely upon information or features that mimic certain linguistic constructions, such as imperatives or politeness, have been demonstrated to be useful.. Textual analysis may not be the only way to determine whether or not someone is spewing hate speech. There is a chance that information gained from other modalities (such as pictures sent along with text messages) might be useful as well. In many situations, the only data sets that may be used to make judgments regarding the overall efficacy of these complicated characteristics are those that are not publicly available and that exclusively cover a specific subtype of hate speech, such as bullying of certain ethnic minority. When it comes to identifying hate speech, there is a need for a uniform data set that can be used to compare characteristics and approaches.

REFERENCES

[1] David M. Blei, Andrew Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022.
[2] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992.Class-based n-gram models of natural language. Computational Linguistics, 18(4):467–479.
[3] Karnan, B., Kuppusamy, A., Latchoumi, T. P., Banerjee, A., Sinha, A., Biswas, A., & Subramanian, A. K. (2022). Multi-response Optimization of Turning Parameters for Cryogenically Treated and Tempered WC–Co Inserts. Journal of The Institution of Engineers (India): Series D, 1-12.
[4] Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2):223–242.
[5] Dasari, M. S., & Mani, V. (2020). Simulation and analysis of three-phase parallel inverter using multicarrier PWM control schemes. SN Applied Sciences, 2(5), 1-10.
[6] Latchoumi, T. P., & Parthiban, L. (2022). Quasi oppositional dragonfly algorithm for load balancing in a cloud computing environment. Wireless Personal Communications, 122(3), 2639-2656.
[7] Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. Social Network Analysis and Mining, 4(1):1–14.
[8] Pavan, V. M., Balamurugan, K., & Latchoumi, T. P. (2021). PLA-Cu reinforced composite filament: Preparation and flexural property printed at different machining conditions. Advanced composite materials.
[9] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71–80, Amsterdam, Netherlands, September. IEEE.
[10] Ravipati, S., Mani, V., & Yarlagadda, S. R. (2021). Efficient Control of Sensorless Hybrid Electric Vehicle Using RBFN Controller. Studies in Informatics and Control, 30(4), 87-97
[11] Banu, J. F., Muneeshwari, P., Raja, K., Suresh, S., Latchoumi, T. P., & Deepan, S. (2022, January). Ontology-Based Image Retrieval by Utilizing Model Annotations and Content. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 300-305). IEEE.
[12]Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web, pages 29–30, New York, NY, USA. ACM.
[13]Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 10(4):215–230.
[14]Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. CoRR, abs/1503.03909.
[15]Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In Marie desJardins and Michael L. Littman, editors, AAAI, pages 1621– 1622, Bellevue, Washington, USA. AAAI Press.
[16] Ravipati, S., Mani, V., & Yarlagadda, S. R. (2021). Efficient Control of Sensorless Hybrid Electric Vehicle Using RBFN Controller. *Studies in Informatics and Control*, *30*(4), 87-97
[17] Dasari, M. S., & Mani, V. (2020). Simulation and analysis of three-phase parallel inverter using multicarrier PWM control schemes. *SN Applied Sciences*, *2*(5), 1-10
[18] C Bhuvaneshwari, A Manjunathan, "Reimbursement of sensor nodes and path optimization", Materials Today: Proceedings, 2021, 45, pp.1547-1551
[19] Roselin Suganthi Jesudoss, Rajeswari Kaleeswaran, Manjunathan Alagarsamy, Dineshkumar Thangaraju, Dinesh Paramathi Mani, Kannadhasan Suriyan, "Comparative study of BER with NOMA system in different fading channels", Bulletin of Electrical Engineering and Informatics, 2022, 11(2), pp. 854–861