IJBPAS, November, 2024, 13(11): 5495-5503

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

MACHINE LEARNING TECHNIQUES FOR DIGITAL HARASSMENT DETECTION ON SOCIAL NETWORKS

RENUGADEVI G*, SASI KALA RANI K, YATHESESWAR KARTHICK R, YASH P A AND VIGNESH M

Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Coimbatore – 641008, India

***Corresponding Author: Dr. G Renugadevi: E Mail: grenugadevi@skcet.ac.in**

Received 24th July 2023; Revised 25th Sept. 2023; Accepted 22nd Dec. 2023; Available online 1st Nov. 2024

https://doi.org/10.31032/IJBPAS/2024/13.11.8428

ABSTRACT

Harassment on the internet is a significant issue that affects both adults and teenagers. Errors like depression and suicide have resulted from it. Social media platforms are being urged more and more to monitor their content. The research that follows builds a model for the detection of cyberbullying in text data using natural language processing and machine learning. It uses information from two different types of cyberbullying: hate speech tweets from Twitter and comments based on personal attacks from Wikipedia forums. Four classifiers and three feature extraction techniques are examined to discover which strategy is most effective. For data from Twitter, the algorithm offers accuracy levels above 90%, and for data from Wikipedia, accuracy levels above 80%.

Keywords: ML, NLP, Feature extraction, SVM, RFC, Precision

I INTRODUCTION

Adult social media users' widespread usage of social media has led to a severe rise in cyberbullying and cyber-aggression. As a result, more and more people have been harmed physically, emotionally, mentally, or both as a result of cyber crime. Communication and content sharing are now easier than ever, but hate speech is also being spread and events based on hatred are being planned via social media sites like Twitter and other forums. The anonymity and portability of such media facilitate the spread of hate speech, which finally results

in hate crimes in a virtual setting where conventional law enforcement is unable to intervene.

Any type of speech that disparages an individual or a group of individuals because of their characteristics, such as their religion, race, sexual orientation, or gender, is referred to as "hate speech.". Recent occurrences, such as the decision by Britain to leave the European Union and the terrorist incidents in Manchester and London, have led to an increase in hate speech in the UK against Muslim and immigrant populations. According to EU surveys and studies, hate speech is becoming more prevalent among young people in the European Economic Area (EEA) region. A survey found that 40% of participants had experienced intimidation or assault because of their ideas, and that 80% of respondents had encountered hate speech

online. Hate speech and criminal activity have increased in the US since Trump's victory. An increasing number of global projects have been created in an effort to study the problem more thoroughly and provide feasible solution.

Automated methods for detecting hate speech online have been developed in the past. The two halves of this method are identifying the traits that these phrases utilize to target a certain group and determining whether material is hate speech or not. The project's latter issue is now being studied due to time constraints. Because hate speech lacks discernible, discriminatory features, as shown by our research of the language present in common datasets, detecting it is more challenging. Deep neural network topologies are recommended as feature extractors because they excel at capturing the meaning of hate speech. The efficacy of these strategies is evaluated using data from social media platforms like Twitter. These numbers show a macro-average F1 improvement of 6 percentage points or a hostile content improvement of 9 percent. In order to undertake a more thorough analysis into cyberbullying, this study attempts to remedy earlier mistakes. With minimal human involvement, this research seeks to construct a cyberbullying detection system.

1. LITERATURE REVIEW

Cyberbullying can be recognized using social network mining tools **[1]**. Application to a real-world instance of cyberbullying in order to identify troll accounts on the social network Twitter **[2]**. Collaboratively identifying cyberbullying activity using Twitter data **[3]** Social network bots can identify cyberbullying based on characteristics of bullying **[4]**. Detecting cyberbullying with deep neural networks **[5]**.

2. Existing System

Cyberbullying is the use of technology to target, harass, threaten, or degrade another person. This online conflict frequently turns into threats against some people in real life. Suicide has been used by certain people. According to a theory put up by Patxi Gal'an-Garc'a *et al*., a troll (someone who engages in cyberbullying) on a social networking site always has a real profile to see how other users react to the fake profile. They suggested using machine learning to find these profiles. The identification method looked at a few profiles that are somewhat related to them. The process involved choosing profiles to examine, gathering data from tweets, choosing attributes to utilize from profiles, and using ML to identify tweet authors. A collaborative detection approach was proposed by Mangaonkar *et al*., where numerous detection nodes are interconnected, each using a different or same algorithm, and data and findings are pooled to produce results. A B-LSTM approach based on concentration was proposed by P. Zhou *et al*.

3. MATERIALS AND METHODS

This project's solution to the challenge of detecting cyberbullying involves categorizing content as either containing or not including the two main types of cyberbullying Twitter hate speech and personal attacks on Wikipedia. The suggested approach to identifying hate speech on Twitter makes use of Support Vector Machine (SVM) and Random Forest Classifier. SVM is essentially used to draw a hyperplane that acts as a boundary between data points in (N)-dimensional space where there are many features. One of the better loss functions for this is the margin value hinge function. If there is no miss classification, which means that our model correctly predicted the class of the data point, we merely need to adjust the gradient based on the regularization arguments in **Figure 1**.

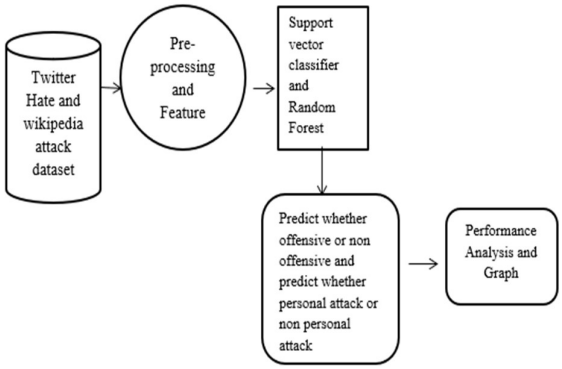

Figure 1: Architecture diagram of the both hate speech detection in twitter and personal attack detection in Wikipedia

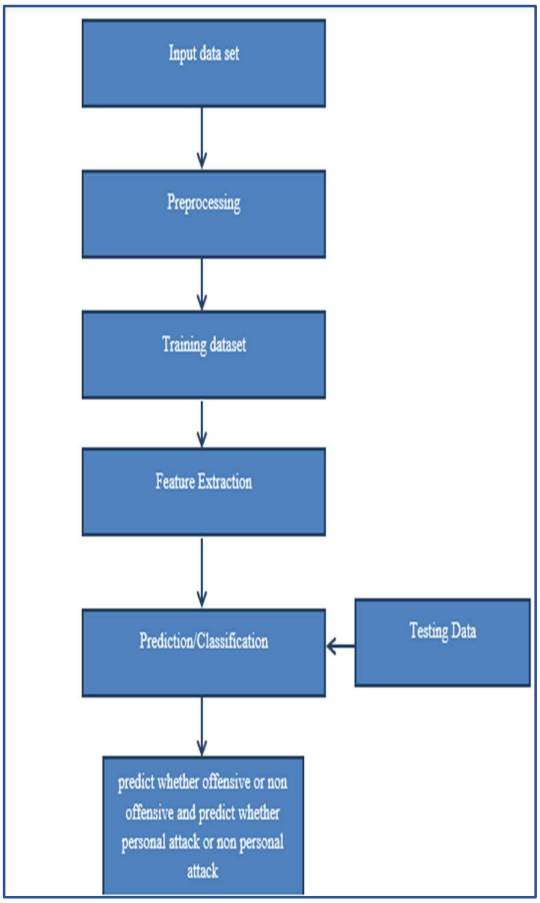

Figure 2: Data flow diagram of the both hate speech detection in Twitter and personal attack detection in Wikipedia

Figure 2 shows activity diagram of the both hate speech detection in Twitter and personal attack detection in Wikipedia.

II RESULTS AND DISCUSSION

1. Data Collection

The actual process of creating a ML model and accumulating data starts now. This stage is critical since the amount and quality of data we can gather will determine how effectively the model performs. Web scraping and other manual interventions are examples of data collection techniques. ML Used to Detect Cyberbullying on social media. The project folder contains the Twitter Hate data set that we provided.31962 distinct pieces of data make up the data set for Twitter Hate Speech Detection. The data set contains three columns, each of which is detailed by using Id: unique id, Labels: B: offensive, A: non-offensive, Tweet: comment

There are 115864 distinct pieces of data in the collection. The data set contains 4 columns, each of which is detailed by

Review Id: unique id,

Comment: comment about Wikipedia titles,

Year: year of comment,

Attack: Personal attack or non-personal

attack

2. Data Readiness

The information will change. by removing any columns and missing data. The titles of the columns that we want to keep or retain will be listed first. The remaining columns are then deleted or discarded, leaving only the columns we want to keep. Finally, we eliminate or remove the rows from the data collection that contain missing values. The steps such as removing unwanted extra symbols, punctuation and stop words need to be done initially and it is subsequently followed by stemming, tokenization, feature extractions, TF-DF vectorize and counter vectorizer with TF-IDF transformer.

3. Feature Extraction

The features in the preprocessed text were then extracted. In the text, there are five characteristics—the rules—to capture hate speech. The final output was a table with the texts listed in the rows and the features stated in the columns, with the sum of all the features being calculated. The method we used to determine how many total rules were included in each tweet in the data set.

4. User Risk Score and Cyberbully Classification

The concept is to view people who deviate from typical behaviours as being riskier, as it is stated throughout the paper. The membership probabilities calculated in the second clustering step really captures these variations. A high membership probability value, to be more precise, denotes that the target user is likely to exhibit one of the behaviours that have emerged from the group to which they belong.

The greatest membership probability value that emerges from the second clustering step is defined as the inverse of the risk score associated with a target user, U. The extracted features will be used to analyze for classification as similar to normal users or risky cyber bully user. Therefore, we set the value of these anomalous features in the test set as randomized value outside the corresponding standard deviation in real world Twitter and Wikipedia data set.

5. RESULT

The SVM, and RFC algorithms were utilized for classification process using training data. The algorithms were processed with the help of the Sci kit-learn Python module. F1 score of the Random Forest Classifier was 80% accurate, while the F1 score of the SVM was same as the RFC. Table I displays the evaluation results of SVM used for twitter hate speech detection while the evaluation findings for RFC for Wikipedia's personal attack detection are shown in **Table 2**.

Table 1: The Performance of the Svm Algorithm

Label	Precision	Recall	F1-Score
Non-personal attack(A)	0.68	0.79	0.73
Personal attack (B)	1.00	0.77	0.87
Accuracy	-	-	0.96
Macro average	0.84	0.78	0.80
Weighted average	0.79	0.77	0.84

Table 2: The Performance of the Rfc Algorithm

Label	Precision	Recall	F1-Score
Non-Hate-speech(A)	0.96	1.00	0.96
Hate speech (B)	0.90	0.50	0.64
Accuracy	-	-	0.99
macro average	0.93	0.75	0.80
Weighted average	0.93	0.79	0.91

Figure 3: Shows the performance of SVM used for Twitter Hate-Speech Detection

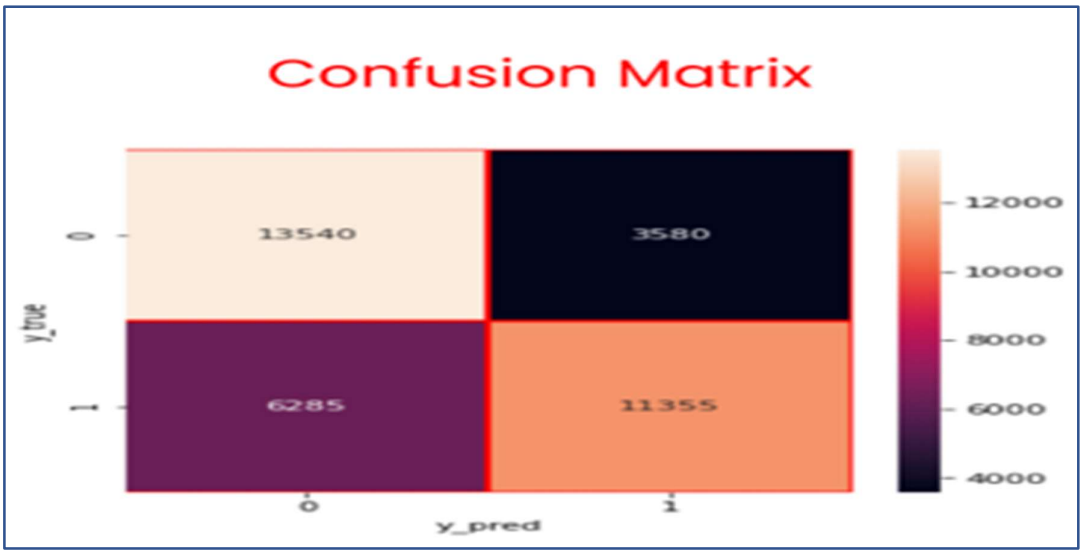

Figure 4: Shows the confusion matrix of SVM used for Twitter Hate-Speech Detection

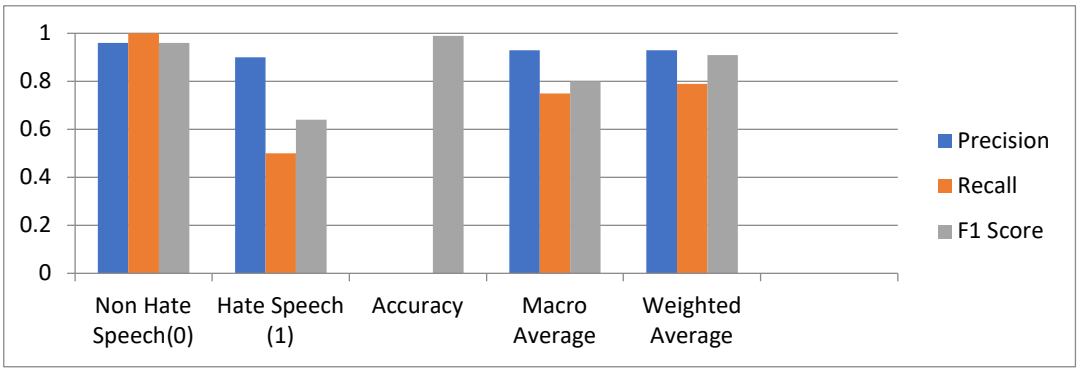

Figure 5: Shows the graphical representation performance of SVM used for Twitter Hate-Speech Detection

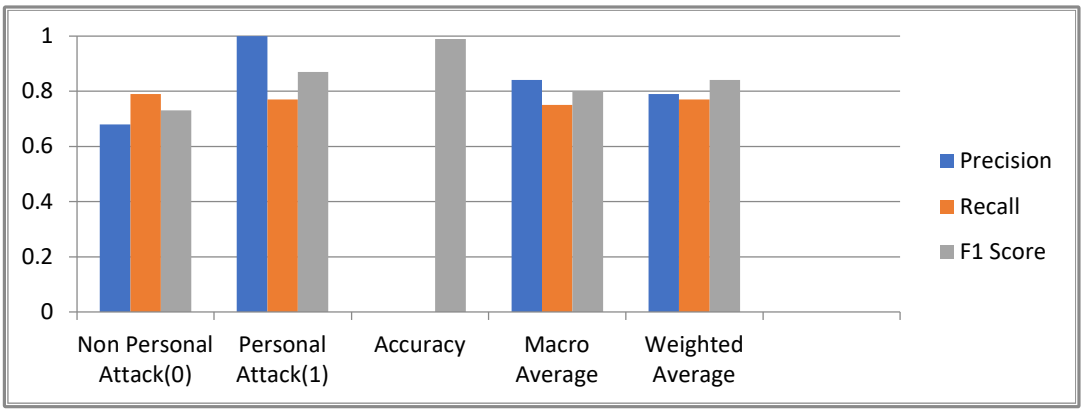

Figure 6: Shows the performance of RFC used for Wikipedia Personal Attack Detection

III. CONCLUSION

There is a need to restrict the growth of cyberbullying because it can be dangerous and result in unfortunate events like suicide, depression, and other problems. Thus, it is crucial to recognize cyberbullying on social media networks. Cyberbullying detection can be used on social media networks to deter users from attempting to engage in such activity once more data and better classified user information for many other types of cyber-attacks are made available. In order to address the issue, we suggested an architecture for cyberbullying detection in this study. Proposed work is personal attacks on Wikipedia and the data architecture for hate speech on Twitter. Natural language processing approaches for this type of speech were successful with accuracy rates of over 90% using basic machine learning algorithms, as hate speech in tweets was often profane, which made it easy to identify.

As a result, BoW and Tf-Idf models outperform Word2Vec models in terms of output quality. Although the three feature selection approaches worked similarly, it was challenging to identify personal assaults using the same model because the comments lacked a lot of learnable sentiment. When integrated with Multi Layered Perceptrons in both datasets, Word2Vec models that take

use of feature context produced comparable performance with significantly less features.

REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018

[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv. 2018.

[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420

[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate

Speech Detection on Twitter," 2016, doi: 10.18653/v1/n16-2013.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.

[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2017, doi: 10.1145/3038912.3052591.

[14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.