Unraveling Twitter Hate Speech: A Comparative Analysis Using LDA and QDA Techniques

Divya Udayan J[1], Veerababu Addanki[2], Nagireddy Moneesh[3], Gandham Sai Ram Pavan [4], Challamalla Satya Srinivas[5]
[12345]Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham, Kollam,Kerala ,India
divyaudayanj@am.amrita.edu[1] *; amenu4aie20004@am.students.amrita.edu*[2]
amenu4aie20150@am.students.amrita.edu[3] *; amenu4aie20125@am.students.amrita.edu*[4] *; amenu4cse21118@am.students.amrita.edu*[5]

***Abstract*—In today's digital age, understanding the tone of online chats, especially tweets, is important, especially tweets, is important. The goal of this study, is to investigate which tweets are of primary concern and which ones are just regular chats. We analyze and detect hate speech from tweets using tools like QDA, LDA. Our findings, helps in facilitating early detection of hate speech which can be used for policy reforms and awareness building in social media.**

***Index Terms*—Twitter Chats, Tweets, QDA, LDA.**

1. Introduction

In the era of digitization and proliferation of online communication platforms, social media has emerged as an essential medium for individuals worldwide to express their views, share news, and communicate with one another. With an estimated 3.96 billion users as of 2022, it is no exaggeration to state that social media shapes public opinion and, in turn, the world in many facets. However, as the digital realm mirrors real life, it is also replete with its challenges. One of the most pressing issues in this domain is the surge of hate speech on platforms such as Twitter.

Hate speech, broadly defined as a public speech that expresses hate or encourages violence towards a person or group based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender, has witnessed an alarming uptick in recent years. It not only affects the mental well-being of individuals but also threatens the very fabric of our democratic societies. Detecting and mitigating the effects of hate speech has, therefore, become crucial. However, manual detection is time-consuming, often inaccurate, and not scalable given the vast volume of content generated on social media. Thus, employing computational methods and leveraging the power of machine learning becomes paramount in this context.

Previous research in the realm of hate speech detection has primarily revolved around traditional textual analysis and basic machine learning models. Moreover, with the dynamic evolution of language on social media platforms, punctuated by the widespread use of slangs, abbreviations, and a unique blend of vernacular languages, the challenge of accurate hate speech detection is accentuated.

Against this backdrop, our research seeks deeper insight into the complexities of hate speech detection on Twitter by using advanced preprocessing techniques and discriminant analysis models. Recognizing the quintessential role of preprocessing in textual data analysis, especially in the noisy and unstructured data from Twitter, our approach aims to enhance the cleanliness and relevance of data. From eliminating Greek symbols, numbers, and certain recurring patterns to replacing slang words with their standard counterparts, our preprocessing techniques are meticulously designed to capture the essence of tweets without the noise. Further, with lemmatization, the data is primed to retain its semantic integrity.

Furthermore, the paper addresses the need for sophisticated models that can encapsulate the intricacies of language while distinguishing between hate and non-hate speech. Our focus on the TF-IDF (Term Frequency-Inverse Document Frequency) approach to convert tweets into a structured format provides a solid foundation for the subsequent use of discriminant analysis. By experimenting with both Quadratic and Linear Discriminant Analysis, we aim to obtain insights into their efficacy in this specific context.

A. Linear Discriminant Analysis(LDA)

Imagine you are at a farmers market, and you see two types of fruits mixed in a basket - apples and oranges. If someone were to ask you to separate them, you'd likely do so by looking at their shape or color. LDA operates on a similar principle. It finds the characteristic or a combination of characteristics (like color or shape for fruits) that best separates two or more classes of objects (or tweets in our case). It tries to find a straight line or plane that best separates the data into these classes.

B. Quadratic Discriminant Analysis(QDA)

Quadratic Discriminant Analysis is similar to the Linear Using the same analogy, let's say some apples are green like oranges, making the earlier method less accurate. This is where QDA comes into play. Instead of trying to draw a straight line, QDA might curve the line, allowing a better fit for those tricky items that do not fall neatly into one category or the other. Essentially, QDA provides more flexibility in separating data by allowing curved boundaries.

The motivation behind our research is manifold. At a macro level, we envision a digital world free from the shackles contribute to the current and understanding. On a technical front, we aspire to current methodologies and offer a robust, scalable solution to a vital problem.Considering Twitter, as a platform known for its brevity and dynamism, our research's outcomes could potentially be adapted to other social media platforms, amplifying its impact.

In summation, this research paper endeavors to advance the field of hate speech detection on social media platforms, with a spotlight on Twitter. By intertwining advanced preprocessing techniques with sophisticated machine learning model. The implications of this research are vast, and its successful execution holds promise for safer, more inclusive digital spaces for all.

2. Literature Survey

The problem of detecting hate speech on Twitter is very necessary. It uses the BERT algorithm which is known for its contextual awareness to detect hate speech. With a 78.69% accuracy, 78.90% precision, 78.69% recall, and 78.77% F1 score for English-language data, the study's testing and analysis has good results. Due to difficulties with the Bahasa Indonesia language it showed a lower accuracy of 68% for Indonesian data. Overall, BERT's success in identifying hate speech which ensures online safety on Twitter[1].The growth of hate speech particularly during important events like elections, is addressed in this research paper, which comes with method detection of hate speech on Twitter using the LSTM approach.The results show that increasing the number of training epochs increases accuracy, with an accuracy rate of 84% when certain parameters like data partition, batch size and learning rate are used. The mode has the ability to lessen the propagation of hatred among Indonesian Twitter users[2].The main aim of the paper is to identify hate speech among different opinions posted by individuals on social sites like twitter.The feature extraction from the data preparation using methods like Porter stemming and Stop Words to improve the quality of the data. The comparison of various machine learning techniques like logistic regression, XGBoost, and random forest have done. Results show random forest has more F1 score, whereas logistic regression and XGBoost classifiers has best validation accuracy. The use of Word2Vec models with skipgram further improves accuracy[3].A recent research introduces a novel approach to data classification using minimum volume enclosing ellipsoids (v-MVEE) instead of conventional techniques like LDA and QDA. The v-MVEE method proves effective in handling outliers, allowing it to disregard abnormal data points. The study compares this approach to traditional methods in cases with limited data and non-standard data collection processes. Results demonstrate that the new approach performs equally well in regular situations and outperforms traditional methods when dealing with outliers or non-standard data collection[4].This study investigates the impact of Covid-19, highlighting the increase in hatred towards Asians on platforms like Twitter. Advanced computer algorithms, specifically Support Vector Machine (SVM) and Random Forest, are employed to analyze and combat this online animosity. The SVM model proves effective in identifying hate-related tweets, with count vectorization being the most successful method. Hate speech peaks are observed in June, July, October, and November, potentially linked to Covid-19 events. The study emphasizes the need to cease using derogatory language on social media and recommends utilizing robust tools to detect hate speech, while also exploring additional effective methods in future research[5].

3. Dataset

The "Twitter Hate Speech" dataset on Kaggle was put together by Rahul Kumar.The size of dataset is 3.16 mega bytes .It's a big collection of tweets, which are short messages from Twitter. Rahul wanted to see which tweets are friendly and which ones are mean or rude. The dataset contains the tweets which are the mixture of the hate and non-hate words. This dataset is like a long table where each line (or row) is a tweet from someone. There are many rows in this table, showing many different tweets. Besides the tweets themselves, the table has several other details (or columns) about each tweet, like if it's mean or not. Rahul's dataset helps people learn about how folks talk to each other online, especially when they're not being nice. By studying these tweets, computer experts can make programs to find and maybe stop rude messages on Twitter. This dataset is a handy tool for those who want to make the internet friendlier for everyone.

4. Methodology

The primary goal is to understand and analyze tweets to determine if they contain hate speech[6]. Through this, we aim to build classifiers that can automatically detect whether a given tweet contains hate speech or not[7].

A. Dataset Information

The dataset is loaded into a DataFrame[8]. Preliminary data information such as the shape, missing values, and the top 5 rows are displayed as shown in Fig. 1

```
Dimention of the data set: (31962, 3)
Tweets with NA value: 0
Labels with NA values: 0
Printing top 10 values of the data set
**************************************
   id  label                                              tweet
0   1      0   @user when a father is dysfunctional and is s...
1   2      0  @user @user thanks for #lyft credit i can't us...
2   3      0                                bihday your majesty
3   4      0  #model   i love u take with u all the time in ...
4   5      0             factsguide: society now    #motivation
```

Fig. 1: Information in the Dataset

B. Preprocessing the data using the NLP techiniques

Next do the preprocessing by applying the NLP techniques on the dataset[9] . Remove the stop words, usernames starting with "@",Strip out numbers, Omit characters like "hm*",

Replace slang words with their proper form using a defined slang dictionary[10] (e.g., ”luv” becomes ”love”).Next apply lemmatization to convert the words to the base form as shown in Fig. 2 .

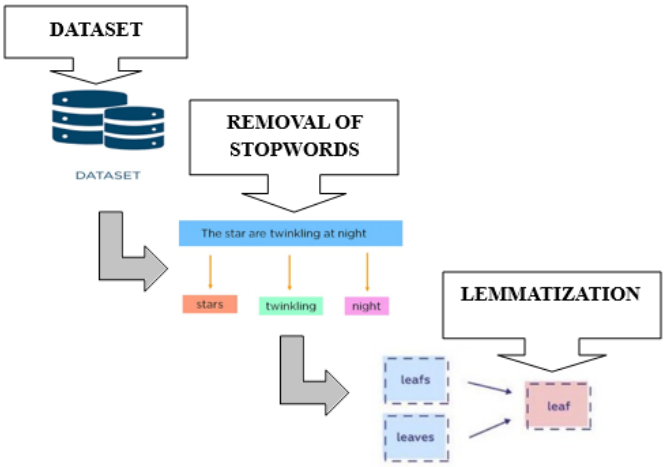

Fig. 2: Processing from Dataset data to Lemmatization

After the lemmatization , tokenization is takes place, in tokenization the suffixes[11] of the word which are common are removed. After the tokenization porter Stemmer is used because this the most powerful[12] technique in the stemming techniques as shown in Fig. 3 .

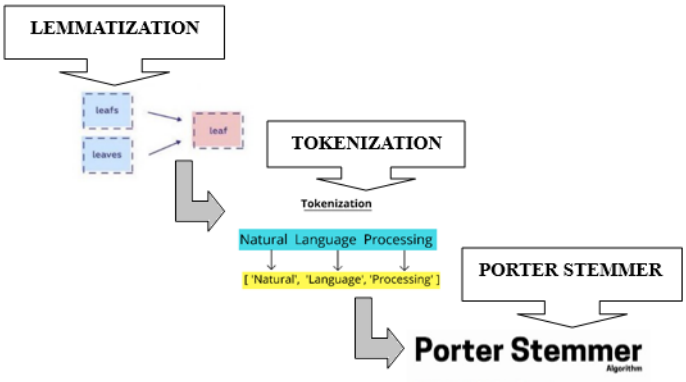

Fig. 3: Processing from Lemmatization to Porter Stemmer

After the porter Stemmer TF-IDF is used to convert the tokenized words into numerical vectors, the TF- IDF approach is used[13]. This method weighs the importance of words across the dataset Fig. 4 .

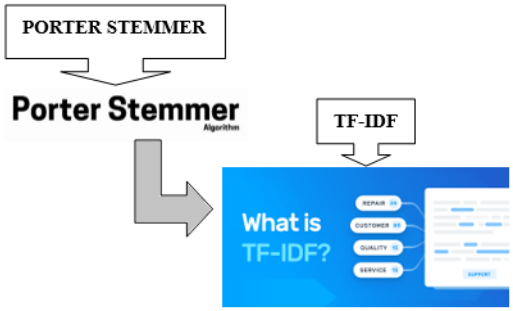

Fig. 4: Processing from Porter Stemmer to TF-IDF

C. Machine Learning Classification Models

Once the data has been preprocessed, the entire dataset is split into testing data and training data[14]. The training data is used for training the classification models and the testing data is used for testing the classification model[15]. 20% of the dataset is used for testing and 80% for training. However, the training data is fed into the QDA and LDA model, i.e. the QDA model and the LDA model are fed into the training data. Once the models are fitted, predict the labels for testing data[16]. After the prediction, evaluate the model by using original testing labels and predicted testing labels, as shown in Fig 5 .

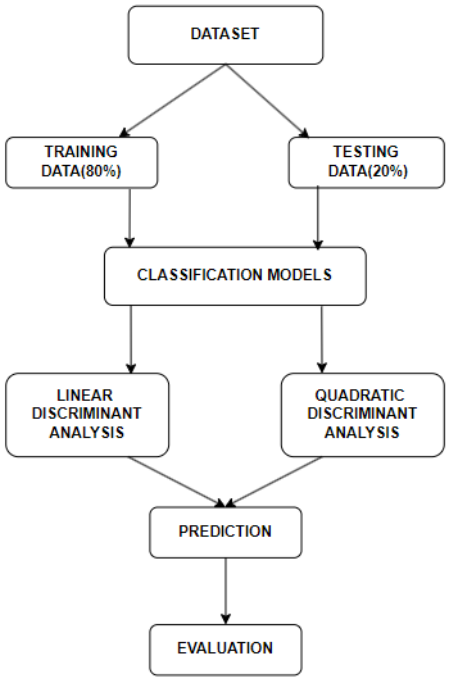

Fig. 5: Processing for Classification

5. Results And Analysis

Using the wordcloud library, two visuals are created to show Frequently used words in non-hate speech tweets. Common words found in hate speech tweets. The dataset contains the tweets which are the mixture of the hate and non-hate words. The Fig. 6 is the word cloud with the width of 800 ,height of 500 and the maximum font size of 110 for the non-hate words .Non-hate words are like which will give the positive vibe while hearing for example friend ,good etc.

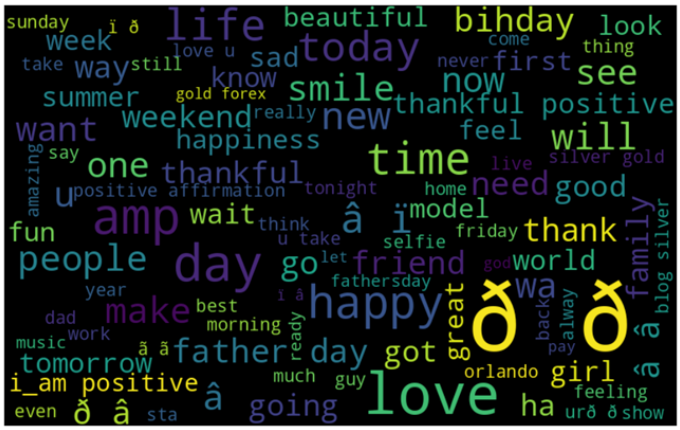

Fig. 6: World Cloud For Positive Hearing

The below Fig. 7 is the word cloud with the width of 800 ,height of 500 and the maximum font size of 110 for the hate words . hate words are like which will give the negative vibe while hearing for example hate, stop, racism etc.

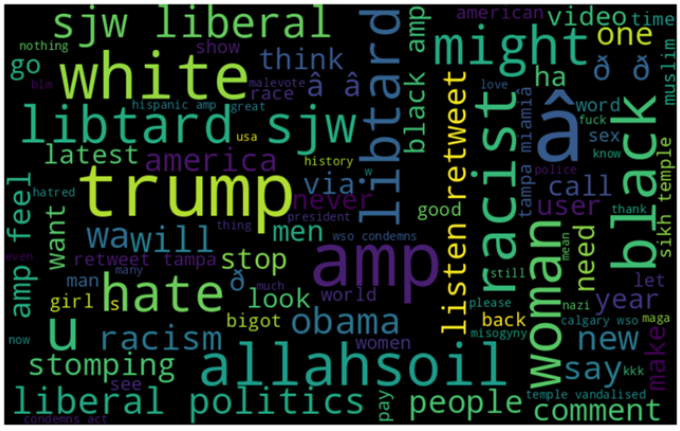

Fig. 7: World Cloud For Negative Hearing

A. Linear Discriminant Analysis

The heatmap for the Linear Discriminant Analysis is shown in Fig. 8. Accuracy is 94 percent from the heat map, precision is 64 percent for Hate ,recall is 44 percent for hate and f1-score is 52 percent for the hate from the heatmap. Precision is 95 percent for non-hate ,recall is 98 percent for non-hate and f1-score is 97 percent for non-hate.

Fig. 8: Heatmap Of LDA

The below Fig. 9 is the Confusion matrix while using the Linear Discriminant Analysis .The Linear Discriminant Analysis model predicted the 5771 non-hates and 218 hates correctly.125 non-hates are wrongly classified as hates and 279 hates are wrongly classified as non-hates.

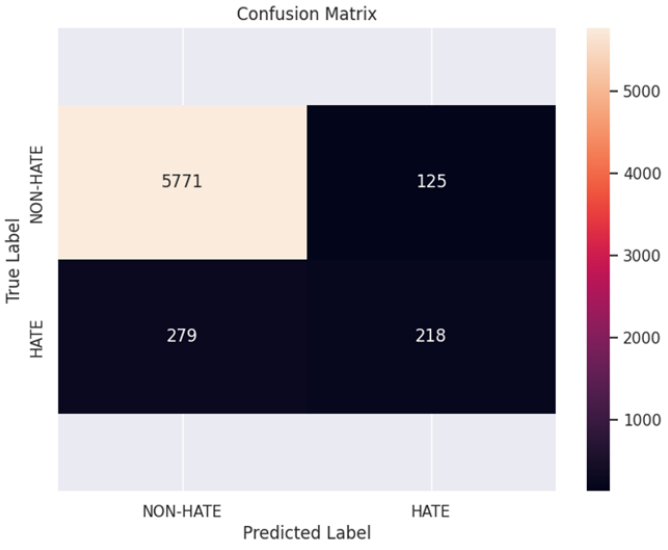

Fig. 9: Confusion Matrix Of LDA

The Fig. 10 shows the ROC curve while using the Linear Discriminant Analysis model, where black line represents the

ROC curve of class Non-hate and green line represents the ROC curve of class hate and x-axis represents the False Positive Rate and y-axis represents the True Positive Rate.

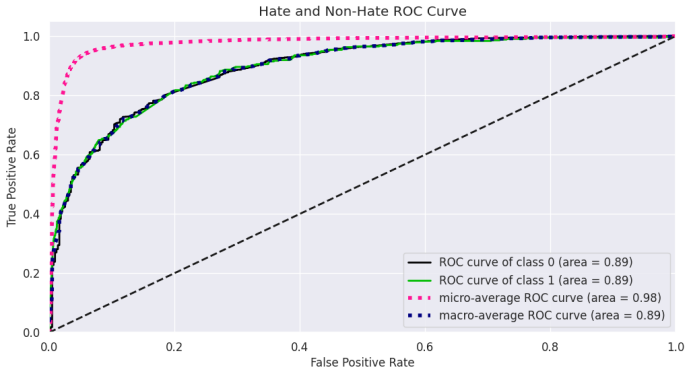

Fig. 10: ROC Curve of LDA

B. *Quadratic Discriminant Analysis*

The heatmap for the Quadratic Discriminant Analysis is shown in Fig. 11. Accuracy is 59 percent from the heat map, precision is 15 percent for Hate ,recall is 94 percent for hate and f1-score is 26 percent for the hate from the heatmap. Precision is 99 percent for non-hate ,recall is 56 percent for non-hate and f1-score is 72 percent for non-hate.

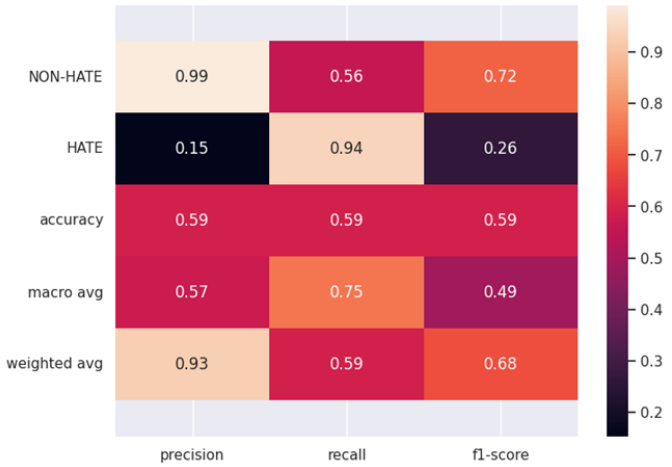

Fig. 11: Heatmap Of QDA

The below Fig. 12 is the Confusion matrix while using the Quadratic Discriminant Analysis. Quadratic Discriminant Analysis model predicted the 3305 non-hates and 468 hates correctly.2591 non-hates are wrongly classified as hates and 29 hates are wrongly classified as non-hates.

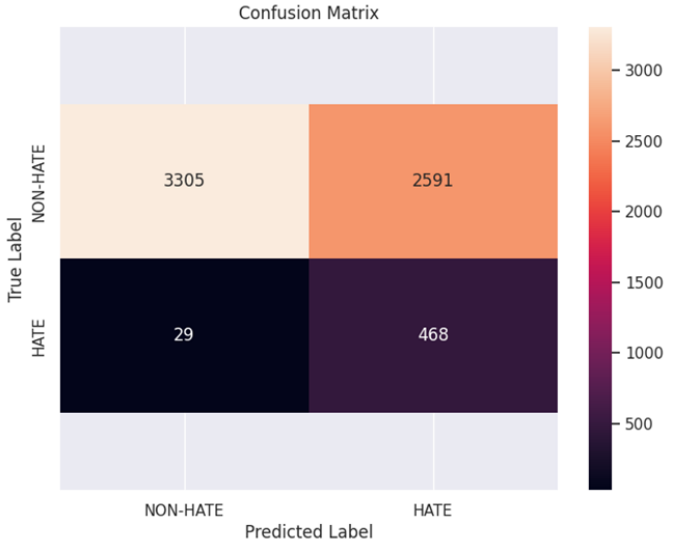

Fig. 12: Confusion Matrix Of QDA

The Fig. 13 shows the ROC curve while using the Quadratic Discriminant Analysis, where black line represents the ROC curve of class Non-hate and green line represents the ROC curve of class hate and x-axis represents the False Positive Rate and y-axis represents the True Positive Rate.

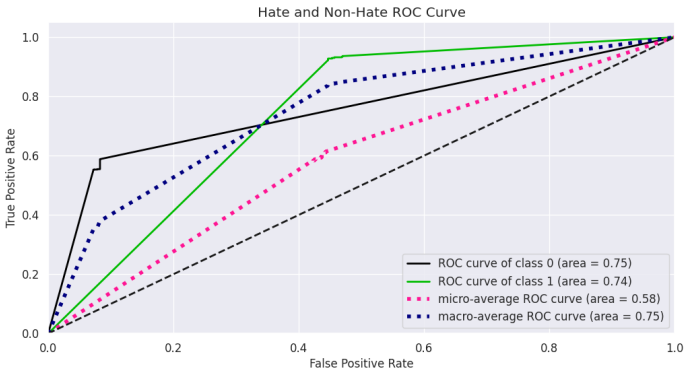

Fig. 13: ROC Curve of QDA

By all above observations we can say that LDA performs better than QDA in this research.

6. Conclusion

The study concludes by comparing classification models like linear discriminant analysis and quadratic discriminant analysis. The performance of the Linear Discriminant Analysis model is much superior to that of the Quadratic Discriminant Analysis model. Less non-hate tweets are predicted as hate

tweets when employing the Linear Discriminant Analysis, and While estimation of non-hate tweets as hate tweets is more common in quadratic discriminant analysis. Therefore, it is better to predict hate tweets as non-hate tweets than vice versa. Important information could be lost if tweets that are not hateful are labelled as such. Compared to the linear discriminant analysis, the quadratic discriminant analysis is less accurate. And the Linear Discriminant Analysis's Precision, Recall, and F1-Sore values were better than the Quadratic Discriminant Analysis.Recognizing hate speech on platforms like Twitter can help improve online safety and user experience. This is a step towards a more respectful digital world, but it is crucial to keep updating our methods as digital communication evolves.

REFERENCES

[1] A. Nayla, C. Setianingsih and B. Dirgantoro, "Hate Speech Detection on Twitter Using BERT Algorithm," 2023 International Conference on Computer Science, Information Technology and Engineeringt (ICCoSITE), Jakarta, Indonesia, 2023, pp. 644-649, doi: 10.1109/ICCoSITE57641.2023.10127831.

[2] S. S. Syam, B. Irawan and C. Setianingsih, "Hate Speech Detection on Twitter Using Long Short-Term Memory (LSTM) Method," 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2019, pp. 305-310, doi: 10.1109/ICITISEE48480.2019.9003992.

[3] A. Razdan and S. S, "Hate Speech Detection using ML algorithms," 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), Gandhinagar, India, 2021, pp. 1-6, doi: 10.1109/AIMV53313.2021.9670987.

[4] P. Juszczak, D. M. J. Tax, S. Verzakov and R. P. W. Duin, "Domain Based LDA and QDA," 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 2006, pp. 788-791, doi: 10.1109/ICPR.2006.461.

[5] S. Shah, X. Yuan and Z. Tyler, "An Analysis of COVID-19 related Twitter Data for Asian Hate Speech Using Machine Learning Algorithms," 2022 1st International Conference on AI in Cybersecurity (ICAIC), Victoria, TX, USA, 2022, pp. 1-6, doi: 10.1109/ICAIC53980.2022.9896967.

[6] E. Krupalija, D. onko and H. Šupić, "Usage of user hate speech index for improving hate speech detection in Twitter posts," 2022 XXVIII International Conference on Information, Communication and Automation Technologies (ICAT), Sarajevo, Bosnia and Herzegovina, 2022, pp. 1-6, doi: 10.1109/ICAT54566.2022.9811159.

[7] A. Tiwari and A. Agrawal, "Comparative Analysis of Different Machine Learning Methods for Hate Speech Recognition in Twitter Text Data," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1016-1020, doi: 10.1109/ICICICT54557.2022.9917752.

[8] A. Kumar, "A Study: Hate Speech and Offensive Language Detection in Textual Data by Using RNN, CNN, LSTM and BERT Model," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1-6, doi: 10.1109/ICICCS53718.2022.9788347.

[9] L. -A. Doan, P. -T. Nguyen, T. -O. Phan and T. -H. Do, "An Implementation of Large Scale Hate Speech Detection System for Streaming Social Media Data," 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Solo, Indonesia, 2022, pp. 155-159, doi: 10.1109/COMNETSAT56033.2022.9994299.

[10] A. Chhabra and D. K. Vishwakarma, "Fuzzy and Machine learning Classifiers for Hate Content Detection: A Comparative Analysis," 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 2022, pp. 1-4, doi: 10.1109/AIST55798.2022.10064822.

[11] K Sreelakshmi, B Premjith, K.P. Soman,Detection of Hate Speech Text in Hindi-English Code-mixed Data,Procedia Computer Science,Volume 171,2020,Pages 737-744,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2020.04.080.

[12] Sreelakshmi K, Premjith B, Gopalakrishnan E A, Soman K P. Hate Speech Detection from Code-mixed Indian Language Using Markov Chains. In Proceedings of the 5th International Conference on Intelligent Computing and Communication (ICICC-2021)

[13] Sreelakshmi K, Premjith B, Soman K P, "Hate speech detection for Hindi-English Code-mixed social media text", in The Seventh International Symposium on Women in Computing and Informatics (WCI'19) 2019.

[14] B. Premjith, Dr. Soman K. P., and Sreelakshmi K., "Amrita CEN at HASOC 2019: Hate Speech Detection in Roman and Devanagiri Scripted Text", in FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, 2019.

[15] M. Venugopalan and Gupta Deepa, an enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis, Knowledge-Based Systems (2022)

[16] J. S. Anjana and Poorna S. S., "Language Identification From Speech Features Using SVM and LDA", 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). Chennai, pp. 1-4, 2018