Hate Speech and Offensive Language Detection Using Ml Algorithms

Dasari sai srujan goud[1], Routhu raviteja[2], Sukka rohith[3] and Dr M V krishna Rao[4]
[123]*Department of CSE(Data Science), Institute of Aeronautical Engineering, Hyderabad*
[4]*Professor, Department of CSE(Data Science), Institute of Aeronautical Engineering, Hyderabad*

***Abstract*—The exponential growth of online social networks (OSNs) and platforms like Twitter, Facebook, and Instagram has made moderating user-generated content, especially hate speech, increasingly challenging. This paper presents a machine learning approach for real-time detection of hate speech on Twitter by leveraging natural language processing (NLP) techniques and machine learning algorithms, including Support Vector Machine (SVM) and Random Forest, to analyze lexical, syntactic, semantic, and contextual features. A comprehensive annotated dataset is used to train the models, evaluated through metrics such as precision, recall, F1 score, and accuracy. The system addresses existing limitations by offering nuanced insights into the context of hate speech, contributing to safer online environments. With enhanced model interpretability and real-time detection capabilities, this scalable solution aims to mitigate the impact of hate speech on social media platforms.**

***Index Terms*—Hate Speech Detection, Machine Learning, Natural Language Processing (NLP), Social Media Moderation, Real-time Detection, Twitter Analysis**

I. INTRODUCTION

Online forums including social media networks and microblogging sites like Facebook, Instagram, and Twitter have developed into vital venues for public conversation in the current digital age. Nevertheless, they have also developed into hubs for derogatory language and hate speech, which poses major threats to internet safety, societal cohesiveness, and individual protection. Because hate speech is subtle and context-dependent, it can be particularly challenging to identify. Hate speech is defined as any communication that denigrates or encourages violence against individuals or groups based on attributes like gender, race, or religion. This project aims to design a very complex system based on improving many strategies that come with hate speech detection. Specifically, the study applies contextual embeddings, linguistic feature analysis, and deep learning and traditional models such as transformers and Support Vector Machines (SVM) for objectionable content detection.

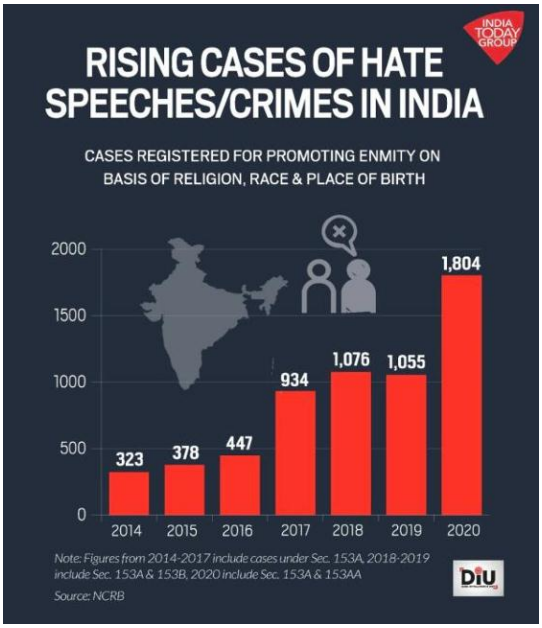

Fig.1. Hate crimes recorded by police (No.of reported hate speech crimes in India)

II. LITERATURE REVIEW

Detection of hate speech and offensive language is perhaps one of the most important issues in today' digital scene, with broad implications for social media applications, online communities, and public discourse as well. This review discusses some key methodologies applied in the recent researches in the field, especially concerning the improvements in detection techniques and current open challenges to be addressed.

techniques and current open challenges to be addressed.

[1] Noted how rule-based algorithms frequently fall short of capturing the subtleties of language, especially when it comes to hate speech. When it

comes to implicit and context dependent expressions of hate, they are ineffectual while recovering explicit statements of hate. False negative rates have increased as a result.

[2] Safari and Zhuang (2020) questioned how emotive and semantic analysis on social media comments could be applied using supervised learning approaches. The present machine learning algorithms—SVMs and decision trees, for example— will improve categorisation more than others, as the authors have shown.

[4] Demonstrated how well CNN and LSTM networks could identify intricate linguistic patterns. The authors claim that deeper models detect more context and semantics than previously thought, which increases the detection rates of subtler hate speech. [16] Reddy (2019), sentiment shifts and linguistic elements in Facebook news posts are influenced by contextual circumstances. As a result, detection frameworks need to account for these shifts. Extensive methods of natural language processing, such BERT contextual embeddings, have been demonstrated by researchers to improve the detection of offensive words.

[24] Expanded on this strategy by utilising machine learning to identify events from Facebook posts in Bengali and Banglish, highlighting the significance of linguistic characteristics in recognising hate speech in a variety of languages and dialects.

[25] Ghiassi and Lee (2018) used supervised machine learning to create a domain-transferable lexicon set for Twitter sentiment analysis. The work demonstrated the possibility of creating domain-invariant lexicons and improved sentiment recognition accuracy, both of which could potentially help identify hate speech on social media sites like Twitter.

[9] Emphasised the effects of linguistic variety on detection accuracy, emphasising the challenges associated with identifying context-dependent hate speech expressions, sarcasm, and irony.

[21] GPT-2 can create conversations without the need for task specific training when used in task-oriented dialogue systems. Their research demonstrated how GPT-2 might lessen the requirement for specialised data while also enhancing discourse quality. The adaptability of pretrained models in conversational AI was illustrated by this method.

III. HATE SPEECH AND OFFENSIVE LANGUAGE DETECTION

Fig. 2 Showshate speech and offensive language detection system analyzes user-generated content from social media platforms and transforms such content into structured data for processing. Natural language processing techniques involved include word embedding and contextual embedding. The system adopts these techniques to capture the linguistic nuances in identifying hate speech. Machine learning models, including SVM and deep learning architectures, classify the content accurately. The variety of detecting methods, then, will help enhance the accuracy and reliability while giving a strong solution to mitigate hate speech online.

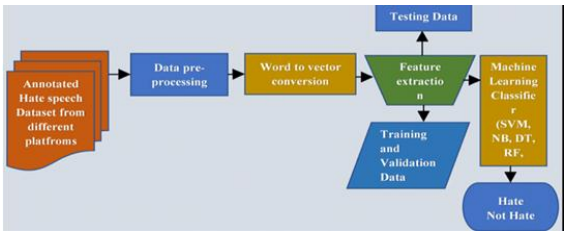

Fig.2 Components of the proposed system

These are the definitions of hate speech and offensive language, with 0 denoting hate or offensive language and 1 denoting non hate or non-offensive language:

• Hate speech and Offensive language (0): Hate speech is defined as any verbal, written, or graphic communication that targets, discriminates against, or encourages violence or other negative actions against any individual or group on the basis of attributes like race, ethnicity, religion, gender, sexual orientation, or other attributes.

• Non-Hate/Non-Offensive Language (1): All of the communication channels that promote civil, productive dialogue and are not founded on hate, discrimination, or violent messages. Positive or neutral, non-malevolent content that doesn't damage individuals or communities. One distinguishing factor between hate speech (zero) and non hate or non-offensive language (1) is that the former is the most effective means of determining the impact that communication has on society. While non-offensive language helps to facilitate validation and

productive discourse in carrying out the aforementioned actions, hate speech positively adds to the continuation, hurt, and disintegration of society. To improve the security and dignity of both the virtual and physical spaces, it is imperative to categorise and identify these kinds of exchanges.

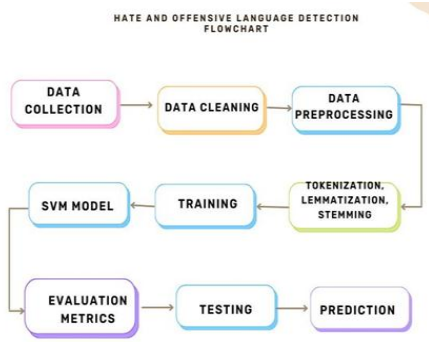

Fig.3 illustrates the work flow of the proposed methodology for developing the hate speech and offensive language detection.

A. Data Collection

7,984 tweets in all were manually gathered from Twitter using search filters that were based on user interactions, hashtags, and keywords. The text, username, date, likes, and retweets of every tweet were meticulously documented in an organized manner. Despite being time-consuming, this approach guaranteed high-quality, objective data for sentiment analysis and trend research.

B. Data Cleaning

Initially, the datasets were examined for missing or inconsistent values. Rows with missing entries were either filled with appropriate statistical measures, such as mean or median, or removed if they were deemed insignificant. This step is essential to maintain the integrity of the analysis and prevent skewed results.

C. Data Preprocessing

•Tokenisation: Dividing text into discrete words or tokens to facilitate word-level pattern analysis by the model.
• Eliminating Special Characters and Punctuation: To concentrate on significant text content, remove special characters, hashtags, URLs, and punctuation.
• Spell Correction and Lemmatization: Eliminating frequently used terms (such as "is," "the," and") that don't significantly contribute to the text's meaning will lower the noise .
• Removing Numbers: Numerical values might be eliminated depending on the situation if they don't help identify hate or objectionable content.
• Handling Slang and Abbreviations: Normalizing internet slang, emojis, and abbreviations (e.g., "u" to "you") to ensure the text is accurately represented.

D. Data Preparation

Before training, the datasets underwent extensive preprocessing to enhance the quality of the input features. This included scaling the features using Term Frequency and Inverse Document Frequency(TFIDF) to standardize the data distribution and mitigate the impact of varying feature scales. Training Process The training phase involved splitting the processed datasets into training and testing subsets. A common practice was to allocate 70% of the data for training and 30% for evaluation. The models were then fitted to the training data, allowing them to learn underlying patterns and relationships inherent in the dataset.

E. Model Training

Model training is a crucial phase in insider threat detection, where we develop predictive models that can accurately identify hate speech and offensive language based on various input features. This process involves several key steps to ensure the robustness and effectiveness of the models employed.

Selection of Algorithms
In our approach, we utilized several machine learning algorithms tailored for specific detection tasks. For hate speech, we opted for SVM, which is particularly suited for identifying outliers in datasets characterized by a significant imbalance between normal and anomalous instances. For signature-based detection, Support Vector Classifier (SVC) was employed, enabling effective classification of user activities based on established access patterns.

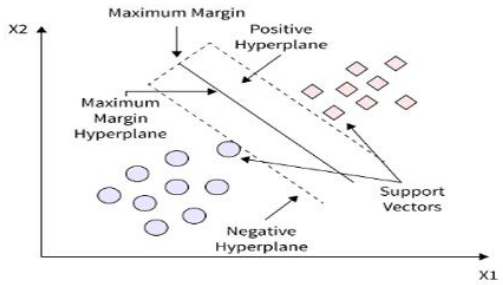

Fig.4 An example of support vector machine

F. Evaluation and Validation

Evaluation is a fundamental aspect of any machine learning project, particularly in the context of insider threat detection, where the stakes are high, and accuracy is paramount. This section outlines the methodologies and metrics employed to assess the performance of our integrated detection framework, focusing on the effectiveness of individual models as well as the overall system. To comprehensively evaluate the performance of our detection models, we utilized several key metrics:

Precision: This metric indicates the proportion of true positive predictions among all positive predictions made by the model. High precision is crucial in reducing false positives, which can lead to unnecessary alerts and resource allocation.

Recall: Recall measures the proportion of true positive predictions relative to the actual number of positive instances in the dataset. A high recall rate is essential for ensuring that as many actual threats as possible are identified, minimizing missed detections. F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. This metric is particularly useful in scenarios where the class distribution is imbalanced, as it combines both metrics into a single value. Accuracy: While accuracy provides a general sense of the model's performance, it is particularly valuable when the class distribution is relatively balanced. However, in the context of anomaly detection, it must be interpreted with caution.

IV. IMPLEMENTATION AND RESULTS

Involving systematic methodologies to validate the effectiveness of the integrated insider threat detection framework. This section outlines the steps taken to design, implement, and evaluate the experiments, ensuring robust and reproducible results.

A. Datasets overview

There are no missing values in the dataset, which has 7,984 rows and 2 columns. It's posts on social media or tweets. There are two columns:

• Tweet: Includes textual material in the form of postings or tweets. It appears that the text has received some preprocessing or augmentation, presumably to enhance its performance for tasks that come next, such as sentiment analysis, categorisation, or pattern recognition

• Label: An numeric value that most likely corresponds to each post's classification label. This label may correlate to categories such as positive/negative sentiment, spam/ham, or another binary distinction, given that the data is binary (0 or 1).

B.Model Selection

• SVM is an effective supervised learning technique that performs exceptionally well in text categorisation, making it particularly useful for identifying hate speech and objectionable language. Finding the optimal hyperplane to divide data into two classes hate speech = 0 and non-hate speech = 1 is the fundamental principle of the SVM method. Support vectors, or the data points that are closest to the decision boundary, are used by SVM to maximise the margin.

• It improves the likelihood that this model will generalise more effectively, as generalisation occurs in a high-dimensional space like text where every word or phrase functions as a feature. SVM is frequently used for dangerous content detection since it does not experience overfitting or It is capable of handling high-dimensional feature sets resulting from the application of text vectorisation techniques like word embeddings and TF-IDF.

C. Experimental Execution

Training the Models: Each model was trained separately on the relevant features extracted from the preprocessed datasets. A standardized training procedure was implemented .This approach helped mitigate overfitting and ensured generalizability across different subsets of data. Parameter Tuning: Hyperparameter optimization was conducted to identify the best-performing configurations for each model. Techniques such as grid search or random search were employed to systematically evaluate different combinations of parameters, optimizing model performance.

D. Performance Evaluation

Evaluation Metrics: Each model's performance was assessed using a comprehensive set of evaluation metrics, including precision, recall, F1-score and accuracy. These metrics provided valuable insights into the models' effectiveness in identifying malicious activities and maintaining a low false positive rate.

• Accuracy: Measure of overall correctness of the model's predictions.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Where,
• TP: True Positives (correctly predicted positives)
• TN: True Negatives (correctly predicted negatives)
• FP: False Positives (incorrectly predicted positives)
• FN: False Negatives (incorrectly predicted negatives)
• Precision: Proportion of correctly predicted insider threats among all predicted threats.

$$Precision = \frac{TP}{TP + FP}$$

• Recall (Sensitivity): Proportion of actual insider threats correctly identified by the model.

$$Recall = \frac{TP}{TP + FN}$$

• F1-score: Harmonic mean of precision and recall, providing a balanced evaluation metric.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Comparative Analysis: The results from the individual models were compared against those of the integrated framework to determine the benefits of combining multiple detection strategies. This analysis involved visualizing the performance metrics and evaluating the trade-offs between different models.

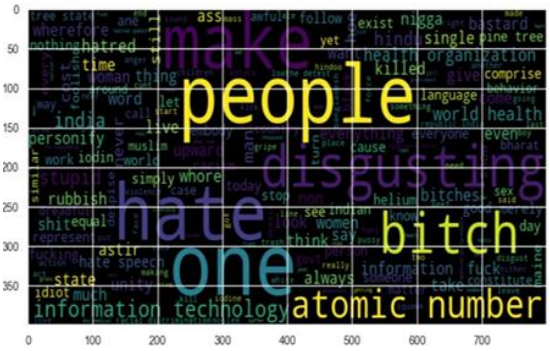

Fig.7 Word cloud of frequency of hate and offensive words.

D. Results

Result Compilation: The outcomes of the experiments were compiled into a structured format, detailing the performance metrics for each model and the integrated system. This compilation allowed for easy comparison and identification of the most effective approaches for insider threat detection. We see the interpretation of each model using the relevant datasets in Fig. 8,We see the results of SVM models.

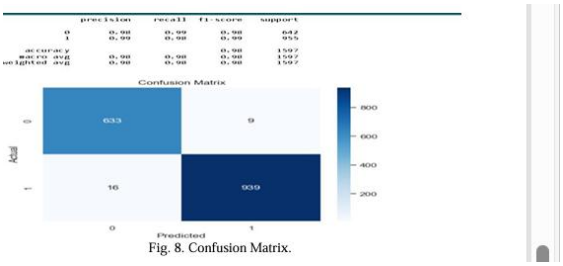

Fig. 8. Confusion Matrix.

V. CONCLUSION AND FUTURE SCOPE

This study conducted the task where the SVM models were applied to social media sites to identify hate speech and objectionable language. The model was able to classify dangerous information with remarkable accuracy through complex natural language processing methods like contextual analysis and word embeddings. The combining of excellent classification powers by SVM with the preprocessing techniques like TF-IDF has heavily enhanced hate speech identification, even on unbalanced datasets. Overall, the system's efforts to counter the exchange of hate speech online have shown much promise, in making digital spaces safer and more welcoming. Future research can investigate how to improve context comprehension in hate speech detection by integrating more intricate deep learning models like transformers.

Furthermore, hate speech in many languages and dialects could be addressed by implementing multi-lingual and cross-lingual models. Enhancing real-time detection systems through low-latency application optimisation of SVM and NLP approaches could be another possible path, allowing for faster and more scalable monitoring on bigger platforms.

ACKNOWLEDGMENT

Required resources are provided by the Department of CSE(DS), Institute of Aeronautical Engineering,

Hyderabad, India for this paper's research study and related work.

REFERENCES

[1] P. Mehta and S. Pandya, ‘‘A review on sentiment analysis methodologies, practices and applications,’’ Int. J. Sci. Technol Res., vol. 4, pp. 601–609, Dec. 2020.

[2] M. N. Moghadasi, Z. Safari, and Y. Zhuang, "A sentimental and semantical analysis on Facebook comments to detect latent patterns," in Proc. IEEE Int. Conf. Big Data (Big Data), Dec. 2020, pp. 4665 4671.

[3] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop on Natural Language Processing for Social Media (SocialNLP), Valencia, Spain, 2017, pp. 1–10.

[4] R. G. Almonte, C. R. Malizon, M. R. F. Montano, and J. Olimpiada, ‘‘Sentiment analysis of local college in the Philippines using Facebook posts towards good governance: A framework proposal,’’ in Proc. 4th Int. Conf. Inf. Comput. Technol. (ICICT), 2021, pp. 47–51.

[5] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolutional neural network," in Proc. 3rd Int. Workshop on Natural Language Processing for Social Media (SocialNLP), 2018, pp. 85–92.

[6] S. Agarwal and A. Sureka, "Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter," in Proc. Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM), 2015, pp. 409–416.

[7] A. Mishra, N. Yannakakis, and L. Togelius, "A survey on deep learning for hate speech detection," IEEE Trans. Affective Comput., vol. 12, no. 1, pp. 3–24, 2021.

[8] A. D. Ward and R. W. Hamm, "Data-driven approaches to hate speech detection on social media platforms: A review," Computers & Security, vol. 101, p. 102027, May 2021.

[9] M. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech detection algorithms: A case study," IEEE Trans. Comput. Soc. Syst., vol. 7, no. 1, pp. 209 218, Mar. 2020.

[10] T. J. Blodgett, L. Green, and D. O’Connor, "Mitigating bias in hate speech detection systems," in Proc. AAAI Conf. on Artificial Intelligence, 2021, pp. 473–481. E

[11] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 85:1–85:30, Jul. 2018.

[12] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in Proc. 25th Int. World Wide Web Conf. (WWW '16), 2016, pp. 145–153.

[13] E. Dinan, V. Logacheva, and X. Li, "Multi-dimensional understanding of abusive language detection using language models," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 4717–4726.

[14] Z. Wang, X. Hu, and J. Hu, "Hate speech detection on social media: A systematic review," IEEE Access, vol. 8, pp. 177868–177883, Oct. 2020.

[15] J. Park and J. Fung, "Detecting online hate speech using neural networks," in Proc. IEEE Int. Conf. Big Data (Big Data), Dec. 2020, pp. 3625–3634.

[16] T. Reddy, ‘‘An analysis of the popularity of Facebook news posts,’’ Projects Appl. Data Sci., vol. 4, p. 1, Oct. 2019.

[17] A. ElSherief, S. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in Proc. 13th Int. AAAI Conf. Web and Social Media (ICWSM '19), 2019, pp. 255–264.

[18] M. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. NAACL Student Research Workshop, 2016, pp. 88–93.

[19] Y. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in Twitter: A comparative study," in Proc. 26th Int. World Wide Web Conf. (WWW '17), 2017, pp. 759–760.

[20] T. Klein and M. Nabi, ‘‘Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds,’’ 2019, arXiv:1911.02365., pp. 1152- 1164, June 2021.

[21] P. Budzianowski and I. Vulià, ‘‘Hello, It’s GPT-2 - how can i help you? Towards the use of pretrained language models for task-

oriented dialogue systems,'' 2019, arXiv:1907.05774.

[22] S. H. An and O. R. Jeong, ''A study on the psychological counseling ai chatbot system based on sentiment analysis,'' J. Inf. Technol. Services, vol. 20, no. 3, pp. 75–86, 2021

[23] S. S. Tu, ''Limits of using artificial intelligence and GPT-3 in patentprosecution,'' SSRN, 2021

[24] N. Dey, M. S. Rahman, M. S. Mredula, A. S. M. S. Hosen, and I.-H. Ra, ''Using machinelearning to detect events on the basis of Bengali and banglish Facebook posts,'' Electronics, vol. 10, no. 19, p. 2367, Sep. 2021.

[25] M. Ghiassi and S. Lee, ''A domain transferable lexicon set for Twitter sentiment analysisusing a supervised machine learning approach,'' Expert Syst. Appl., vol. 106, pp. 197–216, Sep. 2018.

[26] A. S. Kiciman and L. Tran, "A large-scale study of user interactions on social media platforms," IEEE Trans. Comput. Soc. Syst., vol. 8, no. 4, pp. 829–841, Aug. 2021.

[27] Z. Chen, Y. Li, and W. Wang, "Detecting offensive language in online user comments," IEEE Access, vol. 8, pp. 107457–107468, June 2020.