EasyChair Preprint

№ 13799

Hate Speech Analysis and Moderation on Twitter Data Using BERT and Ensemble Techniques

Sukriti Narang, Sejal Karki, Suhani Chauhan, Keshav Garg and Surender Samant

July 3, 2024

Hate Speech Analysis And Moderation On Twitter Data using BERT And Ensemble Techniques

Sukriti Narang[1], Sejal Karki[2], Suhani Chauhan[3] Keshav Garg[4], Surender Singh Samant[5]
Department of Computer Science and Engineering, Graphic Era Deemed to be University, Uttarakhand, India
{sukritinarang2, sejalkarki25, suhanichauhan22002, cskeshugarg86, surender.samant}@gmail.com

***Abstract*—Twitter, a popular social media platform, has become a platform for spreading hate speech, racism, sexism and other sentiments. This has raised ethical, social, and legal concerns, and researchers have developed methods to identify and classify hate speech. This paper investigates Twitter discourse with a focus on detecting hate speech, a prevalent form of online expression. The study utilizes a curated dataset to analyze negative tweets, employing the BERT model and ensemble techniques in the model, trained to detect and classify hateful content. The best classification results were achieved by BERT and CatBoost with hyper-parameter tuning, yielding an accuracy of 92% and 91.1% on the test data, respectively. Additionally, response strategies are devised to moderate content and foster constructive engagement among users. Sentiment analysis is employed to explore the emotional landscape of Twitter discourse. Furthermore, the research is expanded by utilizing clustering to classify hate speech, aiming for a detailed characterisation of online hate speech to enhance our understanding. The analysis encompasses a dedicated exploration of racism and sexism detection, identifying tweets exhibiting bias. The study culminates in providing a comprehensive understanding of online discourse, with potential applications spanning content moderation, user engagement strategies, and the cultivation of a more positive digital space.**

***Index Terms*—BERT, deep learning, ensemble techniques , CatBoost, sentiment , hyper parameter, classification , cluster, moderation**

I. Introduction

The research aims to create a model that accurately classifies tweets as hateful or non-hateful based on hate speech scores, distinguishing between high and mild levels of hate. It also explores the complex realm of racism and sexism within social media discourse, identifying and flagging tweets containing racist content. Sentiment analysis is also explored, categorizing tweets based on their emotional tonality The research's significance extends beyond technical aspects, as understanding and mitigating hate speech , sexism and racism in online spaces is crucial for cultivating a more inclusive and respectful virtual environment. The development of models capable of nuanced hate speech classification and tailored response generation holds promise for creating safer digital spaces that promote healthy dialogue and mutual respect. The research contributes to the ongoing discourse on ethical considerations in the analysis of social media content, fostering a balance between technological innovation and user well-being. This research explore and evaluate diverse advanced machine learning (ML) techniques, with the aim of identifying the most effective approach for detecting and categorizing hate speech. The study undertakes a comparative analysis between the BERT model , a prominent deep learning architecture for its contextual understanding, and advanced ensemble techniques such as CatBoost, AdaBoost, and decision trees [Fig .3]. The primary objective is to ascertain which method yields the highest accuracy in discerning hate speech within text data, particularly on social media platforms like Twitter. By leveraging a combination of these advanced ML techniques, each with its unique strengths and capabilities, the study seeks to optimize the hate speech detection process. Ensemble methods combine multiple models to leverage their collective strengths, ultimately improving overall performance. By aggregating predictions from diverse models such as CatBoost, AdaBoost, and Decision Trees, ensemble techniques offer a powerful approach to mitigating biases and uncertainties inherent in hate speech classification tasks. Through the fusion of complementary models, ensemble techniques effectively capture nuanced patterns in textual data, thereby enhancing the precision and re- call of algorithms. This comparative analysis serves to enhance our comprehension of hate speech detection methodologies. The objective is to refine and strengthen the development of more precise and resilient hate speech moderation tools, thus cultivating a safer online space .The research aims to empower individuals, researchers, and platform administrators to nav- igate the complex landscape of online communication with awareness and efficiency. After classifying the hate tweets sentiment analysis is conducted to detect the intensity of hate, dividing it into low, mild, and high categories. This additional step enhances the understanding of the emotional context surrounding hate speech on Twitter, providing nuanced insights into the varying degrees of negativity expressed within the detected hate speech instances. The hate speech is further categorized into clusters such as racist, sexist, violent, and offensive. This method allows for a more granular understanding of the themes and targets present within the hateful content, enabling a deeper analysis of the underlying motivations and dynamics driving such expressions on Twitter.

II. LITERATURE REVIEW

The transformative impact of social media platforms on communication is evident in the seamless exchange of ideas and opinions across diverse individuals, transcending geographical boundaries. Among these platforms, Twitter stands out as a micro blogging site, condensing expression into 280 characters. While Twitter facilitates global conversations and information dissemination, it also provides fertile ground for

the propagation of hate speech [1] [2], racism, and various sentiments. This literature review navigates through seminal studies to provide a comprehensive understanding of hate speech, racism, and sentiment analysis in the context of Twitter conversations. The spread of hate speech in online communication has become a pressing concern, prompting scholars and technologists to develop robust methods for its identification and classification. Early research, focused on binary classifications, discerning between hateful and non-hateful content [3]. However, this binary approach oversimplifies the intricate nuances of hate speech within the dynamic landscape of online conversations. Recent endeavors, advocate for a nuanced perspective by categorizing hate speech into high and mild levels [4] [5]. This approach recognizes the varying degrees of intensity within hate speech and underscores the importance of tailored responses and communication strategies based on severity. The nuanced classification enriches our understanding of the spectrum of negativity pervading online discourse [6]. The exploration of discriminatory language patterns, particularly with regard to racism, has been a focal point in recent literature conducted a meticulous analysis to identify and flag tweets containing racist content on Twitter. Their work not only sheds light on the prevalence and characteristics of racist and sexist [7] sentiments but also underscores the challenges associated with automated detection of subtle forms of discrimination in user-generated content. Understanding racism in online discourse involves deciphering contextual cues and cultural nuances. According to previous work, tackling racism necessitates a multi-modal approach that considers both text and images [8]. Their work emphasizes the importance of interdisciplinary efforts in combating racism within social media conversations, highlighting the need for more sophisticated models to address the subtleties of racist content. Studies often utilized natural language processing (NLP) techniques to analyze text based data from social media platforms, forums, and other online sources. Researchers experimented with various machine learning algorithms, to classify text as hate speech or non-hate speech. Feature engineering played an important role, with researchers extracting linguistic features such as n-grams, word embedding, and syntactic structures to train models effectively. Early efforts aimed to create a safer online environment by automatically identifying and mitigating hateful content, thereby fostering a more inclusive and respectful digital community. [9] [10] Sentiment analysis has emerged as a pivotal component in decoding the emotional tonality of tweets. Studies have laid the groundwork for categorizing tweets as happy or non-happy, providing valuable insights into the overall sentiment landscape of online communities[11] [12]. Sentiment analysis contributes to a holistic understanding of user emotions, offering a lens through which the collective emotional pulse of online communities can be examined. Beyond binary sentiment classifications, the work introduces more fine-grained sentiment analysis [13], allowing for the recognition of diverse emotional states within text. Integrating such nuanced sentiment analysis with hate speech and racism detection could lead to better understanding of the dynamics within Twitter conversations. As research delves into the complexities of hate speech, racism, and sentiment analysis, ethical considerations become increasingly pivotal. The works have also then explore the ethical implications of hate speech detection models, emphasizing the potential biases and societal impact of automated systems [14]. Understanding the broader societal implications of technology in the context of hate speech and racism detection, [15] emphasize the need for interdisciplinary collaboration to address the ethical challenges posed by automated content moderation [16]. Striking a balance between technological innovation and user well-being is imperative for the responsible advancement of NLP techniques applied to social media content. The literature review lays the groundwork for the research, focusing on hate speech, racism, and sentiment analysis on Twitter. The goal is to deepen understanding and inform policies on moderation and response to hate speech. By addressing nuanced hate speech levels and discriminatory language, the study aim to enhance both academic knowledge and foster responsible online communication. Overall, the review highlights the complex dynamics of hate speech and sentiment analysis, providing a strong basis for the Twitter discourse exploration.

III. ARCHITECTURE

The architecture [Fig. 1] for hate speech detection comprises a multi-layered framework that integrates various components

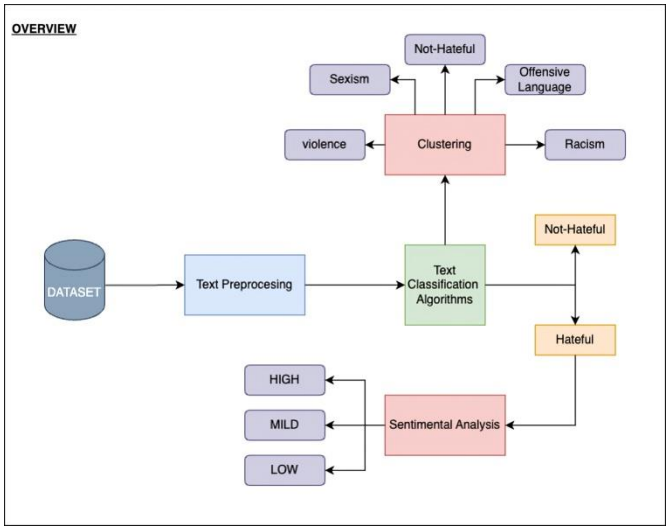

Fig. 1. OVERVIEW OF ARCHITECTURE

to effectively identify and classify instances of hate speech on Twitter. At its foundation, natural language processing techniques preprocess and tokenize textual data, ensuring compatibility with subsequent analysis. These preprocessed features are then fed into machine learning models, including BERT and ensemble techniques like CatBoost, AdaBoost, and Decision Trees, for classification. Sentiment analysis further enriches the understanding by capturing the emotional context of tweets. Additionally, clustering algorithms such as K-Means are employed to categorize hate speech instances and uncover underlying patterns within the data. This comprehensive architecture enables us to gain insights into the prevalence and

the nature of hate speech, empowering us with the knowledge needed to combat online toxicity and foster a safer digital environment.

IV. METHODOLOGY

This section offers a thorough overview of both the problem statement and proposed solution. The methodology begins by detailing the representation of input data and elaborating the approach to modeling sequential steps. Further, delving into the architecture, devised to accomplish this task

A. Data Preprocessing:

The overarching goal was to enhance the quality and cohesiveness of the dataset by eliminating unwanted elements like URLs, special characters, and irrelevant content. This process also involved handling missing data, ensuring text uniformity, and segmenting the text into meaningful units, such as words and phrases, for thorough analysis. This methodology

	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya 
5	3	1	2	0	1	!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! 😂😂😂"
6	3	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"
7	3	0	3	0	1	!!!!“@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!”
8	3	0	3	0	1	" & you might not get ya bitch back & thats that "

Fig. 2. SAMPLE OF THE TWEET DATA

outline the systematic approach employed to clean and standardize text data using the provided preprocess_text() function. The objective is to ensure consistency and enhance the quality of the text data for subsequent analysis:

- **Lowercasing**: All text is converted to lowercase to ensure uniformity and case insensitivity throughout the dataset avoiding discrepancies.
- **Removal of Unwanted Elements:**
  - URLs are eliminated using regular expressions to exclude web links, maintaining the focus on textual content.
  - Twitter mentions (e.g., @username) are removed to prevent user-specific references from influencing the analysis.
  - Hashtags (e.g., #topic) are eliminated to prevent their impact on the analysis.
  - Punctuation marks and special characters are removed to simplify the text and reduce noise.
  - Numeric digits are removed to ensure they don't affect text-based analyses or sentiment assessments.
- **Whitespace Normalization:** Extra white spaces are replaced with single spaces to maintain clean and consistent text formatting eliminating inconsistencies caused by varying spaces and formatting.

This methodology ensures that the text data is appropriately cleaned and standardized, making it suitable for various text-based analyses, including sentiment analysis, classification, and natural language processing tasks. The consistency and quality of the preprocessed text are crucial for obtaining reliable and meaningful insights from the dataset.

B. Natural Language Processing (NLP):

Natural Language Processing (NLP) techniques are used to pre-process the data. The text data was transformed into numerical feature vectors using text-to-vector conversion techniques TF-IDF.

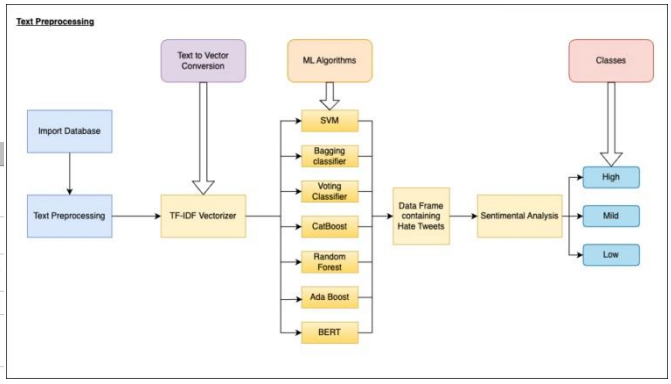

Fig. 3. TEXT PROCESSING FRAMEWORK

'TF-IDF' calculates the occurrence of terms within a document and assesses the importance of terms across a set of documents. It is expressed as follows:

$$\text{TF-IDF} = \frac{\text{Number of occurrences of term in document}}{\text{Total number of terms in document}}$$

$$\text{IDF} = \log \frac{\text{Total number of documents}}{\text{Number of documents containing the term}}$$

The TF-IDF score,

$\text{TF-IDF} = \text{TF} \times \text{IDF}$.

It represents the documents as numeric feature vectors, which can be used with machine-learning algorithms.

C. Classification using ML Algorithms:

Following the preprocessing of raw text data, a range of machine learning techniques are applied, including advanced ensemble methods such as Random Forest, CatBoost, and AdaBoost, alongside the BERT model, as part of the research. Random Forest is a type of ensemble learning technique that involves building multiple decision trees during the training process. These trees are then used to make predictions by combining the most common class (in classification tasks) or

average prediction (in regression tasks) from each tree. AdaBoost is an ensemble learning approach that merges several weak learners to form a robust classifier. It operates through iterative training of weak classifiers on the dataset, with emphasis placed on instances misclassified by preceding classifiers. Finally, it combines the predictions of all weak classifiers through a weighted majority voting scheme to produce a final strong classifier.

CatBoost is a gradient-boosting algorithm, designed for categorical features. It uses a variety of techniques to improve the performance of gradient boosting, such as ordered boosting, pairwise loss function, and gradient-based tree constructions.

A CatBoostClassifier model is initialized with the following parameters:

- `iterations=(100,200,300)`: This specifies the number of trees to train, influencing the model's complexity and ability to capture subtle patterns in the data. A higher number of iterations can lead to better performance but may increase training time.
- `learning_rate =0.05`: This sets the learning rate, determining the magnitude of adjustments made during model training. A lower learning rate can lead to slower convergence but potentially better generalization performance.
- `loss_function='Logloss'`: This defines the loss function used to evaluate the model's performance during training. `Logloss` is a common choice for binary classification tasks, such as identifying negative speech.
- `eval_metric='Accuracy'`: This specifies the evaluation metric utilized to gauge the model's performance on the test set. `Accuracy` serves as a direct measure of the model's capacity to accurately classify tweets as either negative or non-negative.

The model used the training set using the fit() method, enabling it to learn patterns and relationships within the training data that distinguish between negative and neutral language. The model achieved an impressive accuracy of 90% on the test set, demonstrating its effectiveness in identifying instances of negative speech in tweets.

Hyperparameter Tuning:

To further optimize the model's performance, a hyperparameter tuning process is conducted. Different combinations are explored of depth and iterations parameters, which influence
the structure and complexity of the CatBoost trees. This tuning process identifies the optimal combination of hyperparameters that yielded the highest accuracy of 91% on the test set.

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model developed by Google. Leveraging the Transformer architecture and bidirectional training, BERT generates context-aware word embeddings. It can be fine-tuned for various natural language processing tasks, including classification, named entity recognition, and question answering.

BERT model was fine-tuned for sequence classification by adding a single linear classification layer on top, enabling it to predict the sentiment of input sequences. The 'bert-base-uncased' variant is utilized, comprising 12 layers and an uncased vocabulary.

Model Initialization and Configuration:

Upon loading the BERTForSequenceClassification model, configure it with essential parameters:

- `num_labels`: Set to 3, indicating the number of output labels. For binary classification tasks, this would be 2, but it was expanded it to accommodate multi-class classification.
- `output_attentions` and `output_hidden_states`: These were set to False to optimize memory usage and computational efficiency.

Training Process:

To utilize GPU acceleration, the model is transferred to the GPU. Additionally, optimizer is initialized, using AdamW, with a learning rate of 2e-5 and epsilon of 1e-8.A learning rate scheduler is created to adjust the learning rate during training.

The training loop is executed over multiple epochs, with each epoch consisting of iterations over the training dataset. During training, we monitored the loss, accumulated the gradients, and updated the model parameters using backpropagation. The scheduler dynamically adjusted the learning rate, enhancing convergence.

Validation and Performance Evaluation:

After each training epoch, the model's performance is assessed on the validation set. The accuracy is measured as primary evaluation metric, providing insights into the model's ability to correctly classify tweets' sentiment. Throughout the training, track the average training loss and validation accuracy to gauge model performance and convergence.

The training process concluded after the specified number of epochs, resulting in a fine-tuned BERT model capable of accurately classifying tweets. By leveraging BERT's deep contextual understanding of language, coupled with fine-tuning for sequence classification, a commendable accuracy of 92% on the validation set is achieved. This underscores BERT's efficacy in discerning subtle nuances in language and its potential for sentiment analysis tasks in social media discourse.

D. Clustering:

Tweets are categorized and grouped into distinct categories: Sexism, Racism, Violence, and Offensive. Using clustering analysis with the K-means, aims to discern various facets of hate-related information and identify unique clusters representing different aspects of the subject.

K-means is a partitioning clustering algorithm that divides the data into K clusters, where each data point belongs to the cluster with the nearest mean. It iteratively assigns data points to clusters and updates the cluster centroids until convergence. Expanding our research, clustering is utilise to classify hate speech, aiming for a detailed characterization of online

hate speech to enhance our understanding. By leveraging clustering techniques, aimed to identify and categorize various forms of hate speech, enabling a nuanced analysis of the phenomenon and facilitating targeted interventions for mitigating its impact.

E. Sentiment Analysis:

Further the tweets are classified into three categories as low, mild, high depending on the hate score level. A function (categorize_hate_score) is implemented to categorize hate scores into three classes: 'High', 'Mild', and 'Low'. This categorization is based on predefined thresholds.

By categorizing hate scores into three classes, the analysis becomes more granular. This allows for a nuanced understanding of the intensity of hate speech within the dataset.

Each hate score class (Low, Mild, High) corresponds to different levels of severity in hate speech. Providing tailored responses based on the detected severity level enables a more personalized and contextually relevant interaction with users:

- **High:** "We strongly discourage the use of offensive language. Please keep the conversation respectful."
- **Mild:** "Your statement might be perceived as offensive. Let's strive for more positive and respectful communication."
- **Low:** "Your message appears to be neutral and respectful. Keep up the positive tone!"

Social media platforms and communities can use the hate score categorization to manage and moderate content more Effectively high-severity hate speech may require prompt intervention or content removal, while low-severity instances may be addressed through educational initiatives.

Standard Responses: Standard response texts are defined for each hate class. These responses aim to provide constructive feedback and encourage positive communication.

User Guidance: High-severity hate speech often requires a strong and unequivocal response, discouraging the use of offensive language.

Conflict Resolution: Mild-severity hate speech may stem from misunderstandings or unintentional use of offensive language. The response for the 'Mild' category acknowledges the potential perception of offensiveness and encourages users to engage in more respectful communication, promoting conflict resolution.

Positive Reinforcement: For tweets categorized as 'Low,' the response provides positive reinforcement. Acknowledging that the message appears neutral and respectful reinforces positive behavior, fostering a community that values constructive dialogue.

By addressing hate speech promptly and providing appropriate responses, there is a potential to prevent the escalation of conflicts and mitigate the spread of negativity within the online community.

V. PERFORMANCE EVALUATION AND RESULTS

A. Text Classification

Tweets are classified into hate and non-hate categories using a range of ensemble machine-learning algorithms, including AdaBoost , Random Forest, CatBoost , and the BERT model. The testing accuracies of these models are evaluated to assess their effectiveness.

TABLE I
ACCURACY AND PRECISION OF EACH CLASSIFIER

ML Algorithm	Accuracy	Precision	Recall
Bagging Classifier	0.90	0.84	0.89
SVM Classifier	0.90	0.84	0.89
Voting Ensemble Classifier	0.89	0.88	0.84
Random Forest	0.87	0.86	0.68
AdaBoost Classifier	0.90	0.78	0.95
CatBoost Classifier	0.91	0.82	0.92
BERT	0.92	0.91	0.90

[Table. 1] informs us on how the various techniques have performed through various metrics. Among the different models evaluated, CatBoost classifier and BERT emerged as the top performers, achieving the highest accuracy scores of 91.1% and 92%, respectively, showcasing its ability to generalize instances. Hence, meaningful features can be extracted from the textual data, enabling precise classification.

B. Text Clustering

The K-means algorithm was configured to partition the data into five clusters, each representing different categories: racism, sexism, offensive language, violent content, and non-hate speech demonstrated in [Fig. 4]. This categorization enables a more nuanced understanding of the underlying themes and sentiments prevalent in the dataset. The clustering results obtained from K-means offer valuable insights into the distribution of hate speech [Table 2] and elated topics within the dataset, facilitating further analysis and targeted interventions for content moderation and user engagement strategies.

TABLE II
CLUSTERS AND THEIR CATEGORIES FORMED USING K-MEANS

Category	Offensive	Sexism	Racism	Violence
Percentage	36.1336	25.4771	7.8401	0.3914

C. Sentiment Analysis

The sentiment analysis of the twitter data classified as hateful is further categorized into three categories: high, mild and low level of hate. These are with respect to the hate score, i.e. each hate score class (Low, Mild, High) corresponds to different levels of severity in hate speech. Table 3 showcases the hate level analyzed alongside the percentage of tweet from the dataset. Categorizing tweets into severity levels allows for the development of targeted feedback mechanisms.

TABLE III
SENTIMENT ANALYSIS OF HATE SCORE

Hate Level	Percentage of Tweet
High	2.3076
Mild	18.4615
Low	79.2

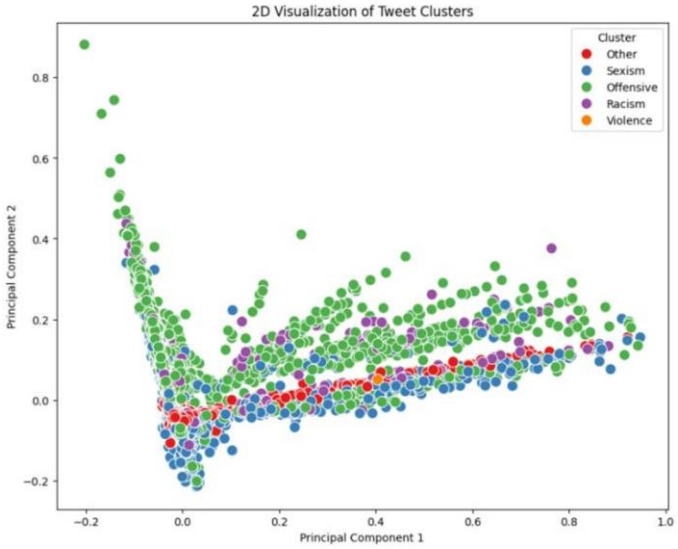

Fig. 4. CLUSTERING OF TWITTER DATA USING K-MEANS

For tweets categorised as 'Low,' positive reinforcement and encouragement can be provided, while tweets categorized as 'High' may trigger more assertive educational responses to discourage offensive language.

VI. Conclusion

In conclusion, the study offers a multifaceted approach to tackle hate speech on Twitter, employing advanced machine learning techniques, sentiment analysis, and clustering algorithms. Through rigorous experimentation, the study demonstrates the efficacy of models like CatBoost and BERT in accurately identifying hate speech instances, achieving impressive accuracies of 91.1% and 92%, respectively. Clustering analysis revealed distinct clusters representing various categories of hate speech, aiding in a nuanced understanding of prevalent themes such as racism, sexism, and offensive language. This insight informs targeted interventions for content moderation and user engagement strategies.

Moreover, the sentiment analysis approach categorizes hate speech instances into low, mild, and high levels based on predefined hate score thresholds, enabling tailored response strategies. By providing positive reinforcement or assertive educational responses, the study aim to mitigate the impact of hate speech and foster a more respectful online environment.

Overall, the research contributes to the ongoing discourse on hate speech detection and moderation, offering practical methodologies and insights for researchers, platform administrators, and policymakers. Moving forward, continued refinement of methodologies and interdisciplinary collaboration will be crucial to address evolving forms of online toxicity while ensuring responsible use of technology

References

[1] Kumar, A. Kumar, S. (2023). Hate Speech Detection in Multi-social Media Using Deep Learning. In: Shaw, R.N., Paprzycki, M., Ghosh, A. (eds) Advanced Communication and Intelligent Systems. ICACIS 2023. Communications in Computer and Information Science, vol 1920. Springer, Cham. https://doi.org/10.1007/978-3-031-45121-86.

[2] H. Ribeiro, M. Calais, P. dos Santos, Y. Almeida, V. Meira Jr, Wagner. (2018). Characterizing and Detecting Hateful Users on Twitter. Proceedings of the International AAAI Conference on Web and Social Media. 12. 10.1609/icwsm.v12i1.15057.

[3] Pereira-Kohatsu JC, Quijano-Sa´nchez L, Liberatore F, Camacho Collados M. Detecting and Monitoring Hate Speech in Twitter. Sensors (Basel). 2019 Oct 26;19(21):4654. doi: 10.3390/s19214654. PMID: 31717760; PMCID: PMC6864473.

[4] Cinelli, M. Pelicon, A., Mozetic˘, I. et al. Dynamics of online hate and misinformation. Sci Rep 11, 22083 (2021). https://doi.org/10.1038/s41598-021-01487-w .

[5] F.Chirigati, "Fighting hate speech and misinformation online", Nat Comput Sci 2, 281–283 (2022). https://doi.org/10.1038/s43588-022-00238-9.

[6] Johnson, N.F., Leahy, R., Restrepo, N.J. et al. Hidden resilience and adaptive dynamics of the global online hate ecology. Nature 573, 261–265 (2019). https://doi.org/10.1038/s41586-019-1494-7.

[7] Lee, E. Rustam, F. Washington, P. Barakaz, Fatima Aljedaani, W. Ashraf, Imran. (2022). Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model. IEEE Access. PP. 1-1. 0.1109/ACCESS.2022.3144266.

[8] C. Bhatt, N. Saini, R. Chauhan and A. K. Sahoo, "Machine Learning Techniques for Hate Speech Detection on Social Media," 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2023, pp. 1-5, doi: 10.1109/CISCT57197.2023.10351228. keywords: Support vector machines;Training;Machine learning algorithms;Social networking (online);Terminology;Hate speech;Supervised learning;Machine Learning;hate speech;twitter;social media;support vector machine,

[9] Shah, S.M.A., Singh, S. (2023). Hate Speech and Offensive Language Detection in Twitter Data Using Machine Learning Classifiers. In:Saini, H.S., Sayal, R., Govardhan, A., Buyya, R. (eds) Innovations in Computer Science and Engineering. ICICSE 2022. Lecture Notes in Networks and Systems, vol 565. Springer, Singapore. https://doi.org/10.1007/978-981-19-7455-7₁7.

[10] G. A. De Souza and M. Da Costa-Abreu, "Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-6, doi: 10.1109/IJCNN48605.2020.9207652. keywords: Twitter;Support vector machines;Feature extraction;Machine learning;Machine learning algorithms;Task analysis;Offensive Language Detection;Naive Bayes;Linear SVM;Attribute Selection;Twitter,

[11] K. K. Pandey, M. Thorat, A. Joshi, S. D, A. Hussein and M. B. Alazzam, "Natural Language Processing for Sentiment Analysis in Social Media Marketing," 2023 3rd International Conference

[12] Nandwani, P., Verma, R. A review on sentiment analysis and emotion detection from text. Soc. Netw. Anal. Min. 11, 81 (2021). https://doi.org/10.1007/s13278-021-00776-6

[13] Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 326-330, doi: 10.1109/ICACITE57410.2023.10182590.

[14] A.Gupta, P. Matta, B. Pant; A comparative study of different sentiment analysis classifiers for cybercrime detection on social media platforms. AIP Conf. Proc. 8 November 2022; 2481 (1): 060005.https://doi.org/10.1063/5.0104639

[15] Su, J., Chen, Q., Wang, Y. et al. Sentence-level sentiment analysis based on supervised gradual machine learning.Sci Rep 13,14500 (2023). https://doi.org/10.1038/s41598-023-414858

[16] Vela´squez, N., Leahy, R., Restrepo, N.J. et al. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. Sci Rep 11, 11549 (2021). https://doi.org/10.1038/s41598-021-89467-