




A comparative analysis of machine learning algorithms for hate speech detection in social media

Esraa Omran ^{1*}

 0000-0002-2227-3478

Estabraq Al Tararwah ²

 0009-0009-4546-7038

Jamal Al Qundus ³

 0000-0002-8848-1632

¹ Center for Applied Mathematics and Bioinformatics, Department of Computer Science, Gulf University for Science and Technology, Kuwait City, KUWAIT

² Gulf University for Science and Technology, Kuwait City, KUWAIT

³ Faculty of Information Technology, Middle East University, Amman, JORDAN

* Corresponding author: husein.i@gust.edu.kw

Citation: Omran, E., Al Tararwah, E., & Al Qundus, J. (2023). A comparative analysis of machine learning algorithms for hate speech detection in social media. *Online Journal of Communication and Media Technologies*, 13(4), e202348. <https://doi.org/10.30935/ojcm/13603>

ARTICLE INFO

Received: 26 May 2023

Accepted: 10 Jul 2023

ABSTRACT

A detecting and mitigating hate speech in social media, particularly on platforms like Twitter, is a crucial task with significant societal impact. This research study presents a comprehensive comparative analysis of machine learning algorithms for hate speech detection, with the primary goal of identifying an optimal algorithmic combination that is simple, easy to implement, efficient, and yields high detection performance. Through meticulous pre-processing and rigorous evaluation, the study explores various algorithms to determine their suitability for hate speech detection. The focus is finding a combination that balances simplicity, ease of implementation, computational efficiency, and strong performance metrics. The findings reveal that the combination of naïve Bayes and decision tree algorithms achieves a high accuracy of 0.887 and an F1-score of 0.885, demonstrating its effectiveness in hate speech detection. This research contributes to identifying a reliable algorithmic combination that meets the criteria of simplicity, ease of implementation, quick processing, and strong performance, providing valuable guidance for researchers and practitioners in hate speech detection in social media. By elucidating the strengths and limitations of various algorithmic combinations, this research enhances the understanding of hate speech detection. It paves the way for developing robust solutions, creating a safer, more inclusive digital environment.

Keywords: hate speech detection, machine learning, social media analysis, text classification

INTRODUCTION

Hate speech, defined as demeaning, discriminatory, or hurtful language directed against individuals or groups based on race, religion, gender, or sexual orientation, has become a significant problem in today's social media platforms (United Nations, 2023). The exponential rise of online communities and the ease with which information may be disseminated have aided the quick spread of hate speech, resulting in negative consequences for people and society. This trend needs effective techniques and technologies to detect and mitigate hate speech, protect online users' safety, and build a more inclusive digital environment.

The prevalence of hate speech in social media is staggering, posing a threat to the principles of tolerance, respect, and equality that underpin democratic societies. Shocking statistics reveal the magnitude of the issue: a recent study found that approximately 40.0% of adults have experienced hate speech or offensive content

on social media platforms through threats of physical violence, sexual harassment, stalking, and more (Laub, 2019). Kindermann (2023) aimed to replace the concept(s) of *hate speech* with *discriminatory speech* by identifying paradigmatic central features namely HATE and GROUP. Following a list of definitions presented by the author, different communities address the concept of hate speech from different perspectives: **from legal to public discourse** *hate speech consists of verbal or non-verbal communication that involves hostility directed towards particular social groups*; **philosophy** *hate speech is a persecutory, hateful and degrading message*; **linguistics** *hate speech is the expression of hatred against persons or groups*; **Oxford English dictionary** *hate speech as a speech or address inciting hatred or intolerance*; **Facebook's community standards** *hate speech as a direct attack against people* (Kindermann, 2023). Moreover, hate speech has been linked to real-world violence and the exacerbation of social tensions, underlining its potential for significant harm and societal disruption.

The consequences of hate speech extend far beyond the online realm. It can contribute to the marginalization and discrimination of vulnerable communities, perpetuating a cycle of hate and prejudice (Kent State University, 2022). Additionally, the viral nature of social media amplifies the impact of hate speech, enabling it to reach a vast audience and potentially incite further acts of hatred and violence. As such, addressing hate speech is an ethical imperative and a matter of public safety and social cohesion.

Efforts to combat hate speech have gained momentum in recent years, with researchers, policymakers, and social media platforms developing algorithms and techniques to identify and address this form of online toxicity. Machine learning has emerged as a promising approach for automated hate speech detection (Sultan et al., 2023). By training models on labeled datasets, machine learning algorithms can learn patterns and linguistic cues associated with hate speech (Yadav et al., 2023a), enabling them to distinguish between offensive content and non-offensive language.

While previous research has explored various methodologies for hate speech detection, there still needs to be more consensus on the most efficient and accurate algorithmic combinations. Individual methods, such as support vector machines (SVM) (e.g., Elzayady et al., 2023) or recurrent neural networks (RNN) (e.g., Mazari & Kheddar, 2023), have been studied in previous research. However, a comprehensive comparative analysis of multiple algorithms is necessary to determine the optimal combination that balances accuracy, simplicity, computational efficiency, and ease of implementation.

This research paper attempts to fill this void by thoroughly examining several machine learning algorithms for hate speech identification in social media. We propose to develop an ideal strategy that offers excellent detection performance while remaining feasible for real-world deployment by exploring the strengths and limits of various algorithmic combinations. This study's findings will not only help to design more effective hate speech detection systems. However, they will also provide vital insights for researchers and practitioners working to promote safer and more inclusive online settings.

LITERATURE REVIEW

The field of hate speech detection in social media has garnered significant attention from researchers aiming to address the challenges associated with identifying and mitigating harmful content. Numerous studies have focused on evaluating different machine-learning algorithms and techniques for effective detection (e.g., Bansal et al., 2022; Das et al., 2023; Simon et al., 2022). This section provides an overview of the existing research in hate speech and spam detection, highlighting key studies that have assessed various algorithms.

Long Short-Term Memory

Long short-term memory (LSTM) was proposed by (Yadav et al., 2023b), as the authors conducted a comparative analysis and assessment of different hate speech detection techniques using machine learning approaches, including deep learning with word embedding.

The results showed that deep learning techniques outperformed other approaches, achieving a performance of over 92.0% using Bi-GRU-GloVe and over 95.0% using LSTM. Bi-GRU is an RNN that processes input data in both forward and backward directions, capturing contextual information from the past and future. Global vectors (GloVe) is a word embedding technique that represents words as dense vector

representations based on their co-occurrence statistics in a corpus. These findings highlight the effectiveness of deep learning algorithms for accurately identifying hate speech, contributing to the advancement of hate speech detection systems.

The research underscores the importance of addressing hate speech on social media and sheds light on the superiority of deep learning techniques. It contributes to the existing body of knowledge by providing insights into the efficacy of automated hate speech detection algorithms. The study is a valuable resource for researchers and practitioners working on detecting hate speech and guiding further advancements in natural language processing (NLP) and social media moderation.

Logistic Regression

Davidson et al. (2017) examined different algorithms for hate speech detection on social media. They aimed to address the challenge of distinguishing hate speech from offensive language. Previous methods relying on lexical detection or supervised learning had limitations in accurately categorizing hate speech.

To overcome these limitations, they employed logistic regression with L2 regularization as their final model for hate speech detection. They used this algorithm to train a multi-class classifier to distinguish between hate speech, offensive language, and non-hateful tweets. Their study showed promising performance, with an overall precision of 0.91, a recall of 0.90, and an F1-score of 0.90.

The authors observed that the model under-classified tweets as less hateful or offensive than human coders. On the other hand, a smaller portion of tweets were mistakenly classified as more offensive or hateful than their correct category. Specifically, approximately 5.0% of offensive tweets and 2.0% of innocuous tweets were incorrectly labeled as hate speech.

Bidirectional Encoder Representations from Transformers

Bidirectional LSTM based deep model explored by Saleh et al. (2023), when combined with domain-specific word embeddings, demonstrated commendable performance in the experiments. The model achieved an impressive 93.0% F1-score when evaluated on a carefully balanced dataset encompassing various hate speech datasets. This outcome underscored the effectiveness of the bidirectional LSTM-based approach, particularly in capturing the nuanced contextual information necessary for identifying hate speech accurately.

In parallel, the study investigated the potential of the bidirectional encoder representations from transformers (BERT) language model in detecting hate speech as a binary classification task. BERT, which is known for its success in various NLP tasks, again demonstrated its high-performance. BERT model achieved an outstanding F1-score of 96.0% on the combined balanced dataset, outperforming the bidirectional LSTM-based deep model.

The findings offer insight into the effect of training data amount on the performance of pre-trained models. Despite the notable difference in corpus size, the bidirectional LSTM-based model with domain-specific word embeddings approached the performance of BERT. This finding underscored the value of utilizing domain-specific data during training, as it provided meaningful insights into the same domain-specific content prevalent in current social media platforms.

METHODOLOGY

CRISP-DM method was applied, and the following main steps were performed in problem understanding, data understanding, analysis, and data preparation (Connolly & Begg, 2005).

Data Collection and Pre-Processing

According to statistics, English is the most commonly used language on social networks (58.8%)¹ and on more than 50.0%² of websites. In addition, "more than half of the literature focuses on figurative language in English" (del Pilar Salas-Zárate et al., 2020). This motivates the language selection for our study.

¹ <https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/>

² <https://www.at-languagesolutions.com/en/idiomas-redes-sociales/>

The data collection process involved obtaining a dataset specifically tailored for hate speech detection from Kaggle, a powerful platform for data science resources—the selected “hate speech and offensive language dataset” comprised 24,783 tweets from diverse Twitter users (Samoshyn, 2020). Rigorous selection criteria were implemented to ensure the integrity and relevance of the data, capturing a wide range of hate speech, offensive language, and related categories prevalent in online discourse.

To prepare the data for analysis, a comprehensive pre-processing pipeline was performed, similar to (Anand et al., 2023). Non-textual information and extraneous metadata were removed, focusing solely on the textual content of the tweets. Text normalization techniques were applied to ensure consistency, such as lowercase conversion, punctuation removal, and handling special characters and emoticons. Stop words, which are ordinary but semantically insignificant, were eliminated to improve efficiency and precision. Advanced linguistic processing techniques, including tokenization and stemming using NLTK library, were employed to segment the text into meaningful units and reduce words to their base or root forms.

The pre-processed dataset was then partitioned into training, validation, and testing subsets, facilitating the systematic development, evaluation, and robust performance assessment of hate speech detection models. Vectorization was performed to convert the textual data into numerical representations suitable for machine learning algorithms, enabling practical analysis, and learning from the data.

By meticulously curating, pre-processing, and vectorizing the dataset, the integrity, relevance, and quality of the data used for training and evaluating the hate speech detection models were ensured. These steps laid the foundation for subsequent analysis and comparative assessment of various machine learning algorithms in effectively detecting hate speech on social media platforms.

Feature Extraction

Feature extraction was pivotal in transforming raw text data into meaningful numerical representations for machine learning algorithms (DeepAI, 2019). This study explored different feature extraction techniques, including n-grams, bag-of-words (BoW), and term frequency-inverse document frequency (TF-IDF), which play essential role for feature extraction (Toktarova et al., 2023).

N-grams were employed to capture different levels of contextual information, ranging from individual words to pairs (bigrams) and triplets (trigrams) of words. By utilizing n-grams of sizes one to three, we encompassed a comprehensive representation of hate speech across diverse forms and contexts.

Additionally, we employed BoW approach, which represents text as a collection of words and their frequencies in the document. This technique disregards the word order but focuses on the presence and frequency of individual words in the text.

Moreover, we utilized TF-IDF method, which assigns weights to words based on their occurrence frequency in a specific document and frequency across the entire corpus. TF-IDF helps to capture the relative importance of words in distinguishing hate speech from other forms of offensive language.

By incorporating these feature extraction techniques into our hate speech detection models, we aimed to capture different aspects of textual information and enhance the models’ ability to accurately identify and classify hate speech.

Algorithm Selection

In the experimentation phase, multiple algorithms were evaluated for hate speech detection. SVM, KNN, random forest, naïve Bayes, and decision tree were considered potential candidates. While some algorithms performed better than naïve Bayes, their computational requirements posed challenges for large-scale datasets, limiting their practical applicability for this research.

The hate speech detection models employing the selected algorithms and features were trained on the training dataset and evaluated using the test dataset. To evaluate the effectiveness of our hate speech detection models, we employed several performance metrics, including accuracy, precision, recall, and F1-score. The comparative analysis with naïve Bayes served as a reference for evaluating the strengths and weaknesses of alternative algorithms and features.

Table 1. Machine learning algorithm comparison

Algorithms	Scores			
	Accuracy	Precision	Recall	F1-score
SVM	0.9061	0.8871	0.9061	0.8862
KNN	0.8277	0.8148	0.8277	0.8078
Random forest	0.8827	0.8692	0.8827	0.8719
Naïve Bayes	0.8616	0.8430	0.8616	0.8322
Decision tree	0.8848	0.8818	0.8848	0.8832
Naïve Bayes + logistic regression	0.8823	0.8677	0.8823	0.8581
Naïve Bayes + decision tree	0.8874	0.8846	0.8874	0.8859

Through this rigorous evaluation process, we aimed to identify the most accurate and efficient approach for hate speech detection, providing valuable insights into the discriminatory power and generalization capabilities of different algorithms and features.

During the experimentation phase, we evaluated multiple algorithms for hate speech detection and ultimately selected naïve Bayes in combination with decision trees as our final model.

Naïve Bayes is a well-established algorithm known for its simplicity and efficiency in text classification tasks (Ray, 2017). It leverages Bayes' theorem and assumes feature independence, making it suitable for distinguishing hate speech from other forms of offensive language.

To enhance the performance of naïve Bayes, we combined it with decision trees, which are versatile and interpretable models that can capture complex relationships between features. Decision trees algorithm complemented the probabilistic nature of naïve Bayes and provided a robust framework for classification.

To assess the performance of the naïve Bayes and decision trees combined model, we partitioned the dataset into training and testing sets. The trained model was utilized to predict the test set labels, categorizing tweets as hate speech or non-hate speech. We compared the predicted and ground truth labels to determine the model's performance using accuracy, precision, recall, and F1-score metrics.

The combination of naïve Bayes and decision trees aimed to leverage the strengths of both algorithms, resulting in a powerful and accurate model for hate speech detection. By integrating naïve Bayes and decision tree algorithms, we achieved enhanced discrimination between hate speech and other forms of offensive language. This integration played a crucial role in advancing hate speech detection techniques.

EXPERIMENTAL RESULTS

The performance of each algorithm was assessed using various evaluation metrics, including accuracy, precision, recall, and F1-score. Additionally, we analyzed the receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) scores for each class (hate speech, offensive language, and non-offensive) (Table 1).

Naïve Bayes and decision trees model exhibited superior performance compared to the other models evaluated in this study. It outperformed KNN, random forest, and naïve Bayes regarding accuracy and F1-score, showcasing its effectiveness in accurately detecting hate speech. Additionally, it demonstrated competitive performance when compared to SVM and other variations of naïve Bayes model combined with other algorithms, such as logistic regression. Naïve Bayes and decision trees model's exceptional accuracy of 0.8874 and F1-score of 0.8859 reinforce its robustness in accurately identifying hate speech instances. Moreover, with a low execution time, it is efficient and suitable for real-time applications. These findings emphasize the model's efficiency, simplicity, and high accuracy, aligning with the criteria sought in hate speech detection.

The successful integration of naïve Bayes and decision trees algorithms provides a promising approach for effectively addressing hate speech on social media platforms. By leveraging the strengths of both algorithms, this model achieves a strong balance between accuracy and computational efficiency, making it a valuable tool for hate speech detection tasks.

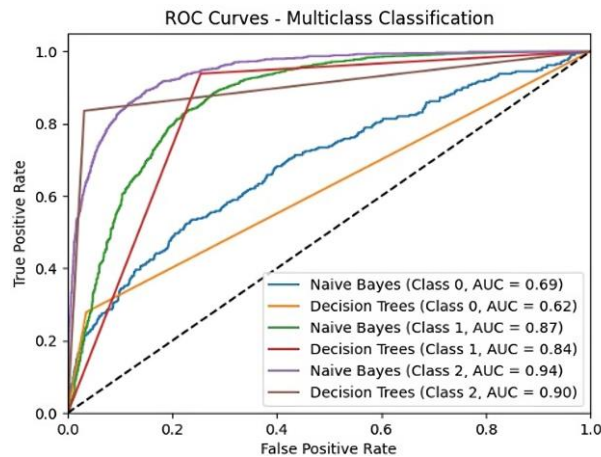


Figure 1. ROC analysis for hate speech, offensive language, & non-offensive tweets classification (Source: Authors)

ROC curve represented in [Figure 1](#) shows the performance of the classifiers (naïve Bayes and decision trees) in distinguishing between different classes of tweets: class 0 (hate speech), class 1 (offensive language), and class 2 (neither).

The x-axis of ROC curve represents the false positive rate (FPR), which is the proportion of tweets belonging to class 2 (neither) incorrectly classified as either hate speech or offensive language. The y-axis represents the true positive rate (TPR), the proportion of correctly classified instances of each class.

The curve shows each class's trade-off between TPR and FPR. The closer the curve is to the top-left corner of the plot, the better the classifier's performance in correctly identifying instances of each class while minimizing false positives.

The plot consists of three curves, each corresponding to one class. The curve for class 0 (hate speech) represents how well the classifiers can distinguish hate speech from the other classes. The curve for class 1 (offensive language) represents the classifier's ability to identify the offensive language. The curve for class 2 (neither) represents the classifier's performance in correctly identifying non-offensive language. AUC summarizes the classifier's performance for each class. A higher AUC indicates better performance distinguishing instances of that class from the rest. The results show that naïve Bayes classifier performs well, with higher AUC scores than the decision trees classifier for all classes.

AUC scores for each class were as follows: class 0 (hate speech)–0.6906 (naïve Bayes) and 0.6218 (decision trees); class 1 (offensive language) –0.8692 (naïve Bayes) and 0.8421 (decision trees); class 2 (neither) –0.9418 (naïve Bayes) and 0.9022 (decision trees). In contrast, accuracy is the difference between correct and incorrect predictions divided by the total number of predictions, so other machine learning classification measures, such as accuracies over 90.0%, provide less meaningful information (Sinyangwe et al., 2023).

These findings underscore the significance of algorithm selection in effectively handling diverse text data. Moreover, the successful integration of ensemble methods highlights their potential in enhancing classification accuracy. ROC analysis further corroborates the models' discriminatory capabilities, as evidenced by AUC scores exceeding 0.5 for all classes, thus confirming their ability to discern between the distinct tweet categories.

In the subsequent section, we will conduct an extensive discussion of these outcomes, exploring the strengths and limitations of our approach and addressing potential implications and avenues for future research.

SYSTEM DESIGN

The system design of our hate speech detection solution encompasses integrating our algorithmic combination to enable real-time detection of hate speech in social media posts. In this section, we will present



Figure 2. Hate speech detection–Instagram post input (Source: Authors, created by Figma)

the design overview and showcase figures that illustrate the usage and functionality of our system. The system incorporates a real-time progress feedback mechanism to provide immediate feedback to the user.

The system continuously analyses and evaluates a message for hate speech content as the user inputs it. Users are informed of the detection progress through a visual representation, such as a progress bar. This real-time feedback ensures that users are promptly aware of the nature of their input.

Figure 2 shows real-time hate speech detection capability of our system. In this scenario, a user is shown interacting with an Instagram³ post, where they input a message that contains hate speech. As the user enters the text, our system processes it using the selected algorithmic combination of naïve Bayes and decision trees. The system then analyzes and evaluates the content against a pre-trained hate speech detection model. The progress bar on top of the post is red, indicating that the message has been identified as hate speech. This immediate detection and visual feedback make users aware of potentially harmful content they are about to share, empowering them to reconsider and refrain from spreading hate speech in real-time.

Figure 3 demonstrates a scenario, where a user intends to post a comment or message that contains hate speech. As the user clicks the post button, our system swiftly analyses the content using the selected hate speech detection algorithms. In real-time, the system detects the presence of hate speech and intervenes to prevent its dissemination. A cautionary popup message immediately appears, notifying the user that their message has been identified as hate speech. The system advises users to reconsider their actions and allows them to revise or refrain from posting the content. This proactive approach is a deterrent, actively discouraging the spread of harmful and offensive language in social media interactions. Our system empowers users to make informed choices by offering timely feedback and guidance, promoting a safer and more respectful online community.

Figure 4 depicts a scenario, where the user inputs a standard text that the system does not detect as hate speech. As the system analyzes the message, the progress bar displayed on top of the interface turns green, indicating that the message is free from hate speech content. This allows the user to proceed with posting the comment without any interruptions. The system design presented in these figures showcases the real-time nature of our hate speech detection solution. By integrating our algorithmic combination into the social media platform, we can analyze user inputs and provide immediate feedback regarding the presence of hate speech. This proactive approach helps prevent the dissemination of harmful content, encourages users to reflect on their online behavior, and fosters a more inclusive and respectful digital environment.

³ This is a demonstration example, the system provided is not limited to Instagram.

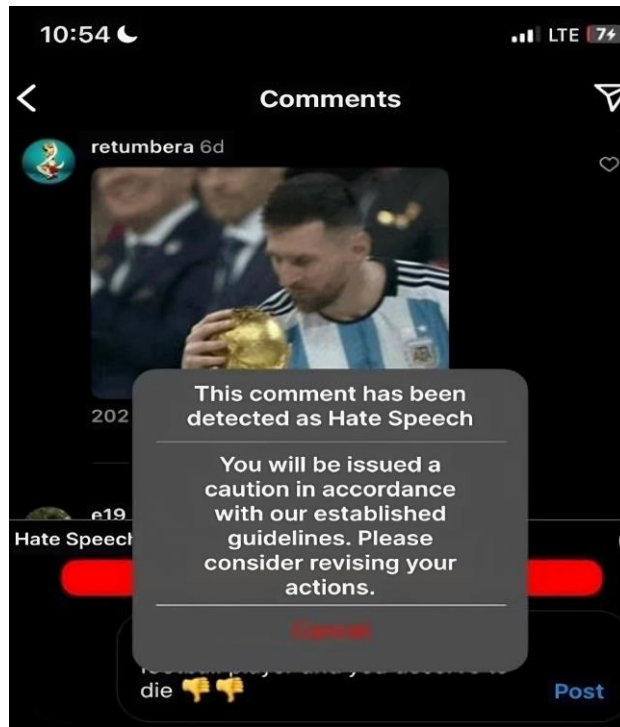


Figure 3. Hate speech detection-caution pop-up (Source: Authors, created by Figma)

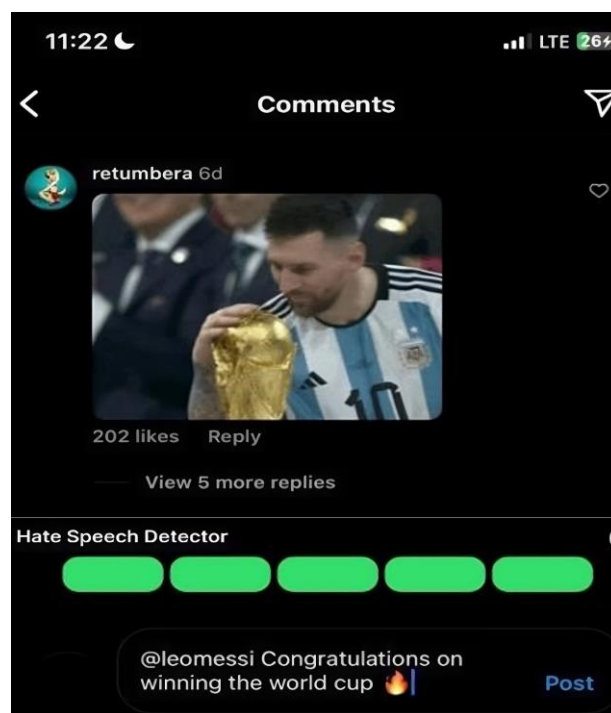


Figure 4. Hate speech detection-normal text input (Source: Authors, created by Figma)

By incorporating these figures into our system design, we effectively demonstrate the functionality and usability of our hate speech detection solution in real-time social media contexts.

DISCUSSION

The study highlights the efficacy of naïve Bayes and decision trees as proficient classifiers for distinguishing between hate speech, offensive language, and non-offensive content within tweets. Naïve Bayes exhibits

promise in accurately classifying non-offensive tweets, whereas decision trees excel in identifying hate speech and offensive language. Voting classifier, which combines the strengths of both models, yields competitive results across all classes. These findings underscore the potential of employing ensemble methods to enhance accuracy in intricate text classification tasks.

Expanding upon these findings, future research should concentrate on several key areas. Firstly, it is imperative to investigate the generalizability of the classification models to diverse datasets and domains. The present study employed a specific dataset, and extending the analysis to varied datasets can offer insights into the robustness and applicability of the models within real-world scenarios. Secondly, developing adaptive models becomes crucial, given the dynamic nature of language and the perpetual evolution of online communication platforms. Exploring the effectiveness of advanced deep learning techniques, such as RNN or transformers, can provide more sophisticated tools for detecting and classifying hate speech and offensive language (Paul & Bora, 2021). Moreover, augmenting the classification models with contextual information and user attributes can enhance their performance. Incorporating user demographics, social network structures, and temporal dynamics can facilitate a comprehensive understanding of the contextual factors associated with hate speech and offensive language. Future research endeavors should aim to investigate these additional features to enhance the accuracy and robustness of the classification models. Lastly, ethical considerations and their potential impact on user privacy and freedom of expression necessitate careful examination (Parker & Ruths, 2023) (e.g., Kebede & Tveiten, 2023; Okpara, 2023). As automated classification systems evolve, ensuring fairness, impartiality, and transparency of the models becomes paramount. Future research should address these ethical challenges by investigating bias detection and mitigation methods, enhancing the models' interpretability, and incorporating diverse perspectives in the training data. Through cautionary popups and informative prompts, users are alerted when their content is flagged as hate speech, allowing them to reconsider their actions. This approach empowers users to make informed choices, fostering a more inclusive and respectful digital environment. The research findings bear significant practical implications as they elucidate the strengths and limitations of diverse algorithmic combinations. Consequently, this study offers valuable guidance for researchers and practitioners engaged in hate speech detection within social media. The identified algorithmic combination satisfies essential criteria such as simplicity, ease of implementation, efficient processing, and robust performance. As a result, it establishes a foundation for developing resilient solutions to foster a safer and more inclusive digital environment. It is important to note that hate speech detection is a complex and evolving field, and there is still room for further advancements and research. While naïve Bayes and decision tree combination showcased excellent performance in this study, future research could further explore the integration of other techniques and algorithms to improve detection accuracy. Additionally, incorporating more diverse and representative datasets would enhance the generalizability of the models.

CONCLUSIONS

The significance of addressing hate speech in social media cannot be overstated, as it threatens the principles of tolerance, respect, and equality in democratic societies. Hate speech perpetuates discrimination and marginalization, inciting real-world violence and exacerbating social tensions. Therefore, developing effective techniques and technologies for hate speech detection is essential for protecting online users' safety and fostering a more inclusive digital environment. The primary goal was to identify an optimal algorithmic combination that is simple, easy to implement, efficient, and yields high detection performance. Various algorithms were explored through meticulous pre-processing and rigorous evaluation to determine their suitability for hate speech detection. Machine learning algorithms, particularly naïve Bayes and decision tree combination identified in this study, have emerged as a promising automated hate speech detection approach. By training models on labeled datasets, these algorithms can learn patterns and linguistic cues associated with hate speech, enabling them to distinguish between offensive content and non-offensive language. The system's design includes intuitive interfaces and proactive measures to promote responsible online behavior.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Ethics declaration: Authors declared that secondary data was used, downloaded from public social media. Authors further declared that the results are ethical, and no ethics committee was required for the study.

Declaration of interest: Authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

- Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., 6 Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. *Theoretical Computer Science*, 943, 203-218. <https://doi.org/10.1016/j.tcs.2022.06.020>
- Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, 3, 100071. <https://doi.org/10.1016/j.dajour.2022.100071>
- Connolly, T. M., & Begg, C. E. (2005). *Database systems: A practical approach to design, implementation, and management*. Pearson Education.
- Das, S., Bhattacharyya, K., & Sarkar, S. (2023). Performance analysis of logistic regression, naïve Bayes, KNN, decision tree, random forest and SVM on hate speech detection from Twitter. *International Research Journal of Innovations in Engineering and Technology*, 7(3), 24-28.
- Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512-515. <https://doi.org/10.1609/icwsm.v11i1.14955>
- DeepAI. (2019). Feature extraction. *DeepAI*. <https://deepai.org/machine-learning-glossary-and-terms/feature-extraction>
- del Pilar Salas-Zárate, M., Alor-Hernández, G., Sánchez-Cervantes, J. L., Paredes-Valverde, M. A., García-Alcaraz, J. L., & Valencia-García, R. (2020). Review of English literature on figurative language applied to social networks. *Knowledge and Information Systems*, 62(6), 2105-2137. <https://doi.org/10.1007/s10115-019-01425-3>
- Elzayady, H., Mohamed, M. S., Badran, K. M., & Salama, G. I. (2023). A hybrid approach based on personality traits for hate speech detection in Arabic social media. *International Journal of Electrical and Computer Engineering*, 13(2), 1979-1988. <https://doi.org/10.11591/ijece.v13i2.pp1979-1988>
- Kebede, S., & Tveiten, O. (2023). Ethnicity as journalism paradigm: Polarization and political parallelism of Ethiopian news in transition. *Online Journal of Communication and Media Technologies*, 13(3), e202335. <https://doi.org/10.30935/ojcm/13333>
- Kent State University. (2022). Negative effects of cyberbullying. *Kent State University*. <https://onlinedegrees.kent.edu/sociology/criminaljustice/community/negative-effects-of-cyberbullying>
- Kindermann, D. (2023). Against 'hate speech'. *Journal of Applied Philosophy*. <https://doi.org/10.1111/japp.12648>
- Laub, Z. (2019). Hate speech on social media: Global comparisons. *Council on Foreign Relations*. <https://www.cfr.org/background/hate-speech-social-media-globalcomparisons>
- Mazari, A. C., & Kheddar, H. (2023). Deep learning-based analysis of Algerian dialect dataset targeted hate speech, offensive language and cyberbullying. *International Journal of Computing and Digital Systems*, 13(1), 965-972. <https://doi.org/10.12785/ijcds/130177>
- Okpara, S. M. N. (2023). Smartphone addiction avoidance via inherent ethical mechanisms and influence on academic performance. *Online Journal of Communication and Media Technologies*, 13(2), e202318. <https://doi.org/10.30935/ojcm/13020>
- Parker, S., & Ruths, D. (2023). Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10), e2209384120. <https://doi.org/10.1073/pnas.2209384120>
- Paul, C., & Bora, P. (2021). Detecting hate speech using deep learning techniques. *International Journal of Advanced Computer Science and Applications*, 12(2). <https://doi.org/10.14569/ijacsa.2021.0120278>
- Ray, S. (2017). Naïve Bayes classifier explained: Applications and practice problems of naïve Bayes classifier. *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2017/09/naive-bayesexplained/>

- Saleh, H., Alhothali, A., & Moria, K. (2023). Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1), 2166719. <https://doi.org/10.1080/08839514.2023.2166719>
- Samoshyn, A. (2020). Hate speech and offensive language dataset. *Kaggle*. <https://www.kaggle.com/datasets/mrmorj/hate-speechand-offensive-language-dataset>
- Simon, H., Baha, B. Y., & Garba, E. J. (2022). Trends in machine learning on automatic detection of hate speech on social media platforms: A systematic review. *FUW Trends in Science & Technology Journal*, 7(1), 001-016.
- Sinyangwe, C., Kunda, D., & Abwino, W. P. (2023). Detecting hate speech and offensive language using machine learning in published online content. *Zambia ICT Journal*, 7(1), 79-84. <https://doi.org/10.33260/zictjournal.v7i1.143>
- Sultan, D., Toktarova, A., Zhumadillayeva, A., Aldeshov, S., Mussiraliyeva, S., Beissenova, G., Tursynbayev, A., Baenova, G., & Imanbayeva, A. (2023). Cyberbullying-related hate speech detection using shallow-to-deep learning. *Computers, Materials & Continua*, 75(1), 2115-2131. <https://doi.org/10.32604/cmc.2023.032993>
- Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylyanbaeva, B., Suieuova, N., & Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. *International Journal of Advanced Computer Science and Applications*, 14(5), 396-406. <https://doi.org/10.14569/IJACSA.2023.0140542>
- United Nations. (2023). What is hate speech? *United Nations*. <https://www.un.org/en/hate-speech/understanding-hate-speech/what-ishate-speech>
- Yadav, A. K., Kumar, M., Kumar, A., Shivani, Kusum, & Yadav, D. (2023a). Hate speech recognition in multilingual text: Hinglish documents. *International Journal of Information Technology*, 15, 1319-1331. <https://doi.org/10.1007/s41870-023-01211-z>
- Yadav, D., Sain, M. K., & Raj B, A. A. (2023b). Comparative analysis and assessment on different hate speech detection learning techniques. *Journal of Algebraic Statistics*, 14(1), 29-48.

