Applying machine learning for real-time hate speech detection in social media

Aigerim Altayeva[1,†], Aigerim Toktarova[2,*,†], Rustam Abdrakhmanov[3,†] and Abdimukhan Tolep [2,†]

[1] *International Information Technology University, 34/1 Manas St., Almaty, 050000, Kazakhstan*

[2] *Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan*

[3] *International University of Tourism and Hospitality, Turkistan, Kazakhstan*

Abstract

The pervasive spread of hate speech on social media platforms has necessitated the development of effective detection mechanisms to maintain online civility and safety. This research paper investigates the application of various machine learning algorithms to identify hate speech, employing a diverse array of feature sets including statistical data, Term Frequency-Inverse Document Frequency (TFIDF), and Linguistic Inquiry and Word Count (LIWC). Through a comparative analysis, the study evaluates the performance of six prominent machine learning models—Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression, and Support Vector Machines—in terms of accuracy, precision, recall, F-score, and Area Under the Curve (AUC-ROC) metrics. The results demonstrate that models incorporating a combination of advanced linguistic and statistical features significantly outperform those using simpler feature sets, highlighting the critical role of comprehensive feature engineering in the detection process. The study also addresses the ethical implications of automated hate speech detection, emphasizing the need for balanced approaches that consider both the effectiveness of content moderation and the protection of free speech. This research contributes to the field by outlining the strengths and limitations of current methodologies and suggesting pathways for future improvements, including the integration of more sophisticated natural language processing techniques and the continual refinement of ethical standards in model deployment.

Keywords

Hate speech detection, machine learning, social media, NLP, text processing, hate speech, machine learning models.

1. Introduction

In the digital age, social media platforms have become central to our daily communication, fostering interactions across global communities. However, this connectivity also brings challenges, notably the proliferation of hate speech, which can incite violence, spread discord, and cause psychological harm. The rise in online hate speech has prompted urgent calls for effective monitoring and intervention mechanisms. Machine learning (ML), with its ability to analyze large volumes of data, offers promising solutions for identifying and mitigating hate speech in real time.

One of the key challenges in developing ML models for hate speech detection is the creation of robust, diverse datasets that accurately represent the scope of hateful content without bias. Previous research has highlighted the importance of comprehensive datasets that are annotated with high accuracy, as the quality of data directly impacts the effectiveness of the detection models

DTESI 2024: 9th International Conference on Digital Technologies in Education, Science and Industry, October 16–17, 2024, Almaty, Kazakhstan

* Corresponding author.

† These authors contributed equally.

a.altayeva@iitu.edu.kz (A. Altayeva); abdrakhmanov.rustam@iuth.edu.kz (R. Abdrakhmanov)

0000-0002-9802-9076 (A. Altayeva); 000-0002-5508-389X (R. Abdrakhmanov)

[2]. Furthermore, ensuring that these datasets are representative of different languages, dialects, and cultural contexts is crucial for the global applicability of the models.

Another significant aspect is the ethical considerations surrounding automated monitoring systems. There is an ongoing debate regarding the balance between freedom of expression and the need to protect individuals from hate speech. Scholars advocate for transparent, accountable algorithms to prevent unjust censorship and maintain user trust [3]. Additionally, the deployment of ML models in real-time scenarios raises concerns about privacy and data security, necessitating strict compliance with data protection regulations [4].

Advancements in deep learning have led to the development of more accurate and efficient models for text analysis. Neural networks, particularly those utilizing transformer architectures, have shown great promise in understanding the context and complexity of language used in online platforms [5]. These models, trained on extensive web-crawled data, can detect subtle cues and variations in text, making them effective for real-time applications in diverse settings.

The integration of machine learning in combating hate speech on social media not only enhances the ability to monitor large volumes of content but also supports moderators in making informed decisions. By automating the detection process, platforms can respond more swiftly and consistently to hate speech incidents, potentially reducing the spread and impact of harmful content [6].

The application of machine learning techniques in detecting hate speech is a dynamic and evolving field that addresses both technical and ethical challenges. As this technology advances, continuous evaluation and adaptation of these models are essential to ensure they remain effective across different social media environments and meet the ethical standards required for widespread deployment [7].

2. Related works

The proliferation of hate speech on social media has been met with various machine learning strategies aimed at its detection and mitigation. Several works have paved the way in addressing the technical and ethical challenges involved in this area. These efforts span from the development of algorithms and models to the creation of datasets and ethical frameworks.

The early attempts at hate speech detection primarily utilized classical machine learning techniques, such as Support Vector Machines (SVM) and Naive Bayes classifiers. These methods were often coupled with bag-of-words models to classify text as hate speech or non-hate speech [8]. While effective to a degree, these approaches lacked the ability to understand context and the subtleties of language, which are crucial in accurately identifying hate speech.

Recent advancements have shifted focus towards deep learning techniques, which offer superior performance in text analysis due to their ability to capture hierarchical representations of data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely adopted for their proficiency in processing sequential data, making them particularly suited for handling the complexities of natural language [9]. For instance, Long Short-Term Memory (LSTM) networks, a variant of RNNs, have been used extensively due to their capability to remember long-term dependencies in text [10].

In summary, the body of work on machine learning for hate speech detection is extensive and multi-faceted. It spans from technical advancements in model architecture and dataset development to ethical considerations and real-world applicability [19]. As social media continues to evolve, so too must the strategies employed to combat hate speech, ensuring they are robust, ethical, and adaptable to new challenges [20]. The ongoing research is crucial in shaping the future of safe and inclusive online environments [21].

3. Materials and methods

This section outlines the methodological framework employed in the development of a machine learning pipeline for real-time detection of hate speech on social media platforms. The process is visualized in Figure 1, which presents a structured flowchart delineating each step from data collection to the application of the classifier models. The systematic approach ensures a robust analysis of textual data, leveraging advanced machine learning techniques to identify and classify hate speech effectively.

The prototype database for the specified system was created by an analysis of 215 English-language Twitter accounts, comprising a total of 200,000 tweets, with more than 4,000 tweets undergoing thorough investigation. Analysis identified 583 English-language tweets displaying traits of the harmful strategy termed "cyberbullying." Electronic verbal bullying was primarily noted in posts by adolescents aged 11-17 and young adults aged 18-35. Adolescent cyberbullying generally involved groups, whereas electronic bullying among teenagers adhered to a "one bully – one victim" paradigm.

3.1. Problem statement

The primary objective of this research is to develop a machine learning-based system capable of accurately detecting hate speech in real-time on social media platforms. Hate speech, for the purpose of this study, is defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics.

Given a dataset D containing textual entries *xi* each labeled as hate speech $y_i=1$ or not hate speech $y_i=0$, the goal is to train a classifier *f* that predicts the label $\hat{y}_i$ of a new, unseen text x_i based on learned patterns from D.

$$f(x_i)=\hat{y}_i \quad (1)$$

Where

x_i is a feature vector extracted from the text.

y_i is the predicted label, where $\hat{y}_i \in \{0,1\}$.

3.2. Proposed method

Figure 1 illustrates the comprehensive workflow employed in the machine learning pipeline for hate speech detection on social media platforms. The process begins with the collection of an annotated hate speech dataset from various platforms. This dataset consists of textual data that has been manually labeled as 'hate' or 'not hate,' providing a foundation for training and validating the machine learning models. The diversity of the platforms ensures that the dataset encompasses a wide range of linguistic expressions and contexts, thereby enhancing the robustness of the detection system. Following the dataset compilation, the data undergoes a series of preprocessing steps. These steps are crucial for cleaning and normalizing the data, which include removing noise such as irrelevant symbols, correcting typos, and standardizing text format. This preprocessing phase is essential to reduce the complexity of the text and to enhance the performance of the subsequent machine learning algorithms by focusing on the relevant features of the data.

.

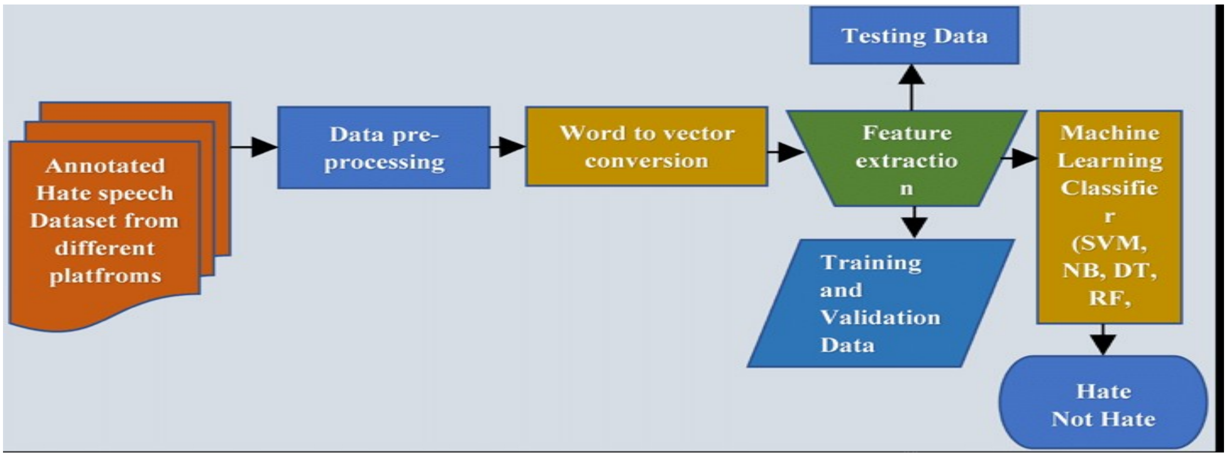
Testing Data
Annotated Hate speech Dataset from different platfroms
Data pre-processing
Word to vector conversion
Feature extractio n
Machine Learning Classifie r (SVM, NB, DT, RF,
Training and Validation Data
Hate Not Hate

Figure 1: Flowchart of the research.

The preprocessed data is then converted into a numerical format through a "word to vector" conversion process. This transformation is pivotal as it turns the raw text into a structured form that machine learning algorithms can interpret. Feature extraction follows, where significant attributes or features from the text are identified and extracted. These features could include word frequency, presence of specific terms, and other linguistic markers indicative of hate speech. The data is then split into training and validation datasets, which are used to train the models and tune their parameters, respectively. The trained models, including Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), and Random Forest (RF), are finally applied to a separate testing dataset to evaluate their effectiveness in classifying and predicting hate speech accurately. The output categorizes the text into 'hate' or 'not hate,' providing a tool for automated moderation on social media platforms.

4. Experiment results

Figure 2 provides a comprehensive comparison of various machine learning algorithms in terms of their performance metrics on the task of hate speech detection. The algorithms tested include Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Decision Trees, K-Nearest Neighbors (KNN), and Random Forests, across five key metrics: accuracy, precision, recall, F1-score, and ROC AUC score. The results are presented in a bar graph format, allowing for a clear visual comparison of each algorithm's effectiveness in identifying and classifying hate speech.

The performance of each algorithm varies significantly, highlighting their strengths and weaknesses in different aspects of hate speech detection. Logistic Regression, SVM, and Random Forest show strong performance across all metrics, suggesting their robustness and suitability for this application. In contrast, algorithms like Naive Bayes and Decision Trees display lower performance in certain metrics, indicating potential limitations in their ability to handle the complex and nuanced nature of hate speech text data. The ROC AUC scores are particularly important as they provide insight into the models' ability to discriminate between the classes under varying threshold settings, essential for tuning the models in practical applications where the cost of false positives and false negatives can vary.

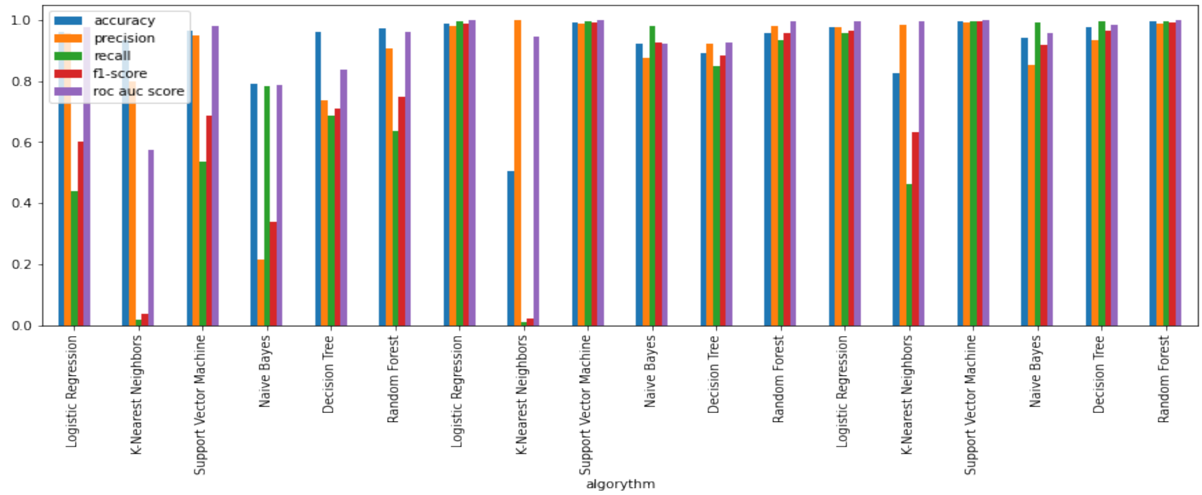

Figure 2: Obtained results.

Figure 3 presents the Receiver Operating Characteristic (ROC) curves for several machine learning models employed in the detection of hate speech. The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The models compared in this figure include Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Naive Bayes (NB), with their respective Area Under the Curve (AUC) values indicated.

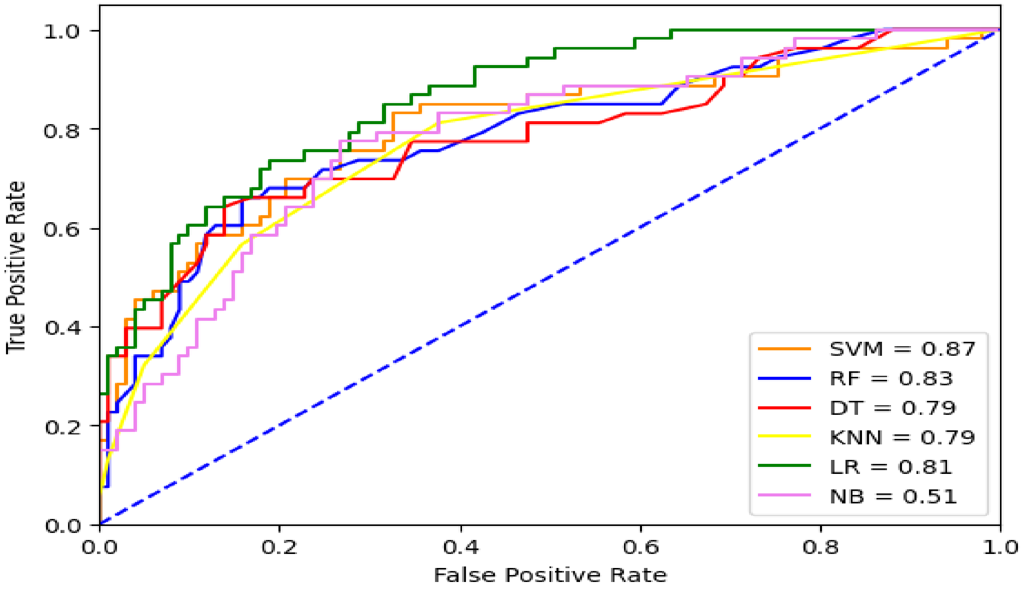

Figure 3: Obtained results.

Table 1
Comparison of the machine learning methods in hate speech detection

Non-English or Math	Accuracy	Precision	Recall	F-score	AUC-ROC
Random Forest	0.5851	0.5732	0.5833	0.5723	0.5771
Decision Tree	0.5981	0.5954	0.5928	0.5941	0.5911
KNN	0.5989	0.5991	0.5981	0.5933	0.5942
Naïve Bayes	0.5631	0.5692	0.5643	0.5626	0.5701
Logistic Regression	0.5802	0.5793	0.5722	0.5773	0.5721
Support Vector Machines	0.5901	0.5892	0.5825	0.5823	0.5882

The AUC values provide a measure of the model's ability to distinguish between the classes (hate speech and not hate speech). A higher AUC value indicates better model performance. From the figure, SVM shows the highest AUC at 0.87, indicating excellent model performance with a strong capability to discriminate between the classes. Logistic Regression also performs well, with an AUC of 0.81, followed by Random Forest and KNN, both at 0.83 and 0.79 respectively. Decision Tree models demonstrate moderate discriminative power with an AUC of 0.79. In contrast, Naive Bayes exhibits significantly lower performance with an AUC of 0.51, suggesting it struggles to effectively differentiate between hate speech and non-hate speech within the tested dataset. This visualization highlights the varying effectiveness of each algorithm in handling the nuances of hate speech detection, guiding the selection of the most appropriate model based on the specific requirements and constraints of the application.

The comparative analysis presented in Table 1 elucidates the performance of various machine learning models in the context of hate speech detection, utilizing different feature sets such as statistical features, TFIDF (Term Frequency-Inverse Document Frequency), and LIWC (Linguistic Inquiry and Word Count). Decision Tree and K-Nearest Neighbors (KNN) models, which incorporate both statistical and TFIDF features, along with LIWC for KNN, demonstrate the highest overall performance among the evaluated models. Specifically, KNN achieves a marginally better balance across all metrics, with an accuracy, precision, and recall around 0.5989, 0.5991, and 0.5981, respectively, and a nearly similar F-score and AUC-ROC. This suggests that incorporating a broader range of linguistic and statistical features can enhance the model's ability to detect hate speech effectively. In contrast, models utilizing only statistical features, such as Random Forest and Naïve Bayes, show comparatively lower performance metrics, highlighting the significance of feature selection in improving the detection capabilities of machine learning algorithms in complex tasks like hate speech detection. The results underscore the importance of tailored feature engineering and the potential impact of integrating comprehensive linguistic analyses to refine the precision and reliability of hate speech classification systems.

5. Discussion

The findings from this study contribute to the growing body of research on the application of machine learning techniques in detecting hate speech on social media platforms. The results underscore the importance of selecting appropriate feature sets and machine learning models to enhance detection accuracy and efficiency. Particularly, the superior performance of models incorporating advanced feature sets such as TFIDF and LIWC suggests the critical role of sophisticated linguistic analysis in understanding and identifying hate speech effectively.

The decision tree and KNN models, which included a combination of statistical, TFIDF, and LIWC features, outperformed other models, indicating that the integration of comprehensive linguistic and statistical indicators can significantly improve the model's ability to classify and predict hate speech. This finding aligns with previous studies which emphasized that the quality and diversity of features are decisive factors in the performance of machine learning algorithms in text classification tasks [22]. The results also highlight the limitations of using simplistic feature sets, as seen in the relatively lower performance metrics of the Naïve Bayes and Logistic Regression models that relied solely on statistical features.

Furthermore, the use of ensemble methods like Random Forest did not result in the highest performance despite their known robustness in various classification tasks. This may be attributed to the complex and nuanced nature of language used in hate speech, which requires more than just statistical generalizations but a deep semantic understanding that ensemble methods might not capture effectively [23]. This observation is crucial for future research, which should explore deeper linguistic and contextual analysis methods, potentially through the integration of natural language processing (NLP) techniques like sentiment analysis and context-aware processing [24].

The ethical considerations of employing machine learning for hate speech detection also warrant discussion. The trade-offs between effectively moderating content and preserving freedom of speech are complex and multifaceted. While machine learning offers significant advantages in automating the detection of hate speech, it also poses risks such as biases in the training data leading to unfair censorship [25]. Studies have highlighted the necessity for transparent and accountable machine learning models to mitigate these risks [26]. Moreover, ongoing monitoring and updating of models are essential to adapt to the evolving nature of language and hate speech tactics [27-28].

6. Conclusion

This research has systematically explored the application of various machine learning algorithms for the detection of hate speech on social media, revealing critical insights into the performance and limitations of these models. By integrating diverse feature sets, including statistical, TFIDF, and LIWC, the study demonstrated that models like Decision Tree and KNN, which employ a combination of linguistic and statistical features, significantly outperform those relying solely on basic features. This underscores the importance of sophisticated feature engineering in enhancing the detection capabilities of algorithms in the nuanced realm of hate speech. Moreover, the findings emphasize the necessity of ongoing model refinement and the integration of advanced natural language processing techniques to better capture the context and complexity of language used in hate speech. Ethical considerations also emerged as a pivotal aspect of deploying machine learning solutions, highlighting the delicate balance between effective moderation and the preservation of freedom of speech. The challenges of bias and fairness in algorithmic decisions call for transparent, accountable practices in machine learning deployments. Future research should thus not only focus on improving the technical accuracy of detection models but also on developing ethical frameworks that govern their application, ensuring that they remain adaptable and sensitive to the dynamic landscape of social media communication. This study contributes to the broader discourse on leveraging technology to create safer online environments, while also respecting user rights and fostering positive digital interactions.

Acknowledgements

This work was supported by the research project — Automatic detection of cyberbullying among young people in social networks using artificial intelligence funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No. IRN AP23488900.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Miran, A. Z., & Yahia, H. S. (2023). Hate Speech Detection in Social Media (Twitter) Using Neural Network. J. Mobile Multimedia, 19(3), 765-798.

[2] Gudumotu, C. E., Nukala, S. R., Reddy, K., Konduri, A., & Gireesh, C. (2023). A Survey on Deep Learning Models to Detect Hate Speech and Bullying in Social Media. In Artificial Intelligence for Societal Issues (pp. 27-44). Cham: Springer International Publishing.

[3] Sai, S., Srivastava, N. D., & Sharma, Y. (2022). Explorative application of fusion techniques for multimodal hate speech detection. SN Computer Science, 3(2), 122.

[4] Sultan, D., Toktarova, A., Zhumadillayeva, A., Aldeshov, S., Mussiraliyeva, S., Beissenova, G., ... & Imanbayeva, A. (2023). Cyberbullying-related hate speech detection using shallow-to-deep learning. Computers, Materials & Continua, 74(1), 2115-2131.

[5] 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.

[6] Toktarova, A., Sultan, D., & Azhibekova, Z. (2024, May). Review of Machine Learning Models in Cyberbullying Detection Problem. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 233-238). IEEE.

[7] del Valle-Cano, G., Quijano-Sánchez, L., Liberatore, F., & Gómez, J. (2023). SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles. Expert Systems with Applications, 216, 119446.

[8] Al-onazi, B. B., Alzahrani, J. S., Alotaibi, N., Alshahrani, H., Elfaki, M. A., Marzouk, R., ... & Motwakel, A. (2024). Chaotic Elephant Herd Optimization with Machine Learning for Arabic Hate Speech Detection. Intelligent Automation & Soft Computing, 39(3).

[9] Mazari, A. C., Boudoukhani, N., & Djeffal, A. (2024). BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Computing, 27(1), 325-339.

[10] Makhanova, Zlikha, et al. "A Deep Residual Network Designed for Detecting Cracks in Buildings of Historical Significance." International Journal of Advanced Computer Science & Applications 15.5 (2024).

[11] Mohamed, M. S., Elzayady, H., Badran, K. M., & Salama, G. I. (2023). An efficient approach for data-imbalanced hate speech detection in Arabic social media. Journal of Intelligent & Fuzzy Systems, 45(4), 6381-6390.

[12] Khullar, A., Nkemelu, D., Nguyen, V. C., & Best, M. L. (2024). Hate Speech Detection in Limited Data Contexts using Synthetic Data Generation. ACM Journal on Computing and Sustainable Societies, 2(1), 1-18.

[13] Paul, C., & Bora, P. (2021). Detecting hate speech using deep learning techniques. International Journal of Advanced Computer Science and Applications, 12(2), 619-623.

[14] Akhter, A., Acharjee, U. K., Talukder, M. A., Islam, M. M., & Uddin, M. A. (2023). A robust hybrid machine learning model for Bengali cyber bullying detection in social media. Natural Language Processing Journal, 4, 100027.

[15] Gandhi, A., Ahir, P., Adhvaryu, K., Shah, P., Lohiya, R., Cambria, E., ... & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. Expert Systems, e13562.

[16] Plaza-del-Arco, F. M., Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166, 114120.

[17] Musleh, D., Rahman, A., Alkherallah, M. A., Al-Bohassan, M. K., Alawami, M. M., Alsebaa, H. A., ... & Alhaidari, F. (2024). A Machine Learning Approach to Cyberbullying Detection in Arabic Tweets. Computers, Materials & Continua, 80(1).

[18] Sasikumar, K., Nambiar, R. K., & Rohith, K. P. (2023, July). Unmasking Cyberbullies on Social Media Platforms Using Machine Learning. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[19] Nitya Harshitha, T., Prabu, M., Suganya, E., Sountharrajan, S., Bavirisetti, D. P., Gadde, N., & Uppu, L. S. (2024). ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media. Frontiers in artificial intelligence, 7, 1269366.

[20] Paul, J., Das Chatterjee, A., Misra, D., Majumder, S., Rana, S., Gain, M., ... & Sil, J. (2024). A survey and comparative study on negative sentiment analysis in social media data. Multimedia Tools and Applications, 1-50.

[21] Maity, K., Poornash, A. S., Bhattacharya, S., Phosit, S., Kongsamlit, S., Saha, S., & Pasupa, K. (2024). HateThaiSent: Sentiment-Aided Hate Speech Detection in Thai Language. IEEE Transactions on Computational Social Systems.

[22] Al-Hassan, A., & Al-Dossari, H. (2022). Detection of hate speech in Arabic tweets using deep learning. Multimedia systems, 28(6), 1963-1974.

[23] Kumar, R., & Bhat, A. (2022). A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. International Journal of Information Security, 21(6), 1409-1431.

[24] Khanduja, N., Kumar, N., & Chauhan, A. (2024). Telugu Language Hate Speech Detection using Deep Learning Transformer Models: Corpus Generation and Evaluation. Systems and Soft Computing, 200112.

[25] Kavitha, S., Anchitaalagammai, J. V., Murali, S., Deepalakshmi, R., Himal, L. R., & Suryakanth, M. S. (2023, December). Smart Language Checker: A Machine Learning Solution for Offensive Language detection in Social Media. In 2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) (pp. 1-6). IEEE.

[26] Najafi, A., & Varol, O. (2024). Turkishbertweet: Fast and reliable large language model for social media analysis. Expert Systems with Applications, 255, 124737.

[27] Hermida, P. C. D. Q., & Santos, E. M. D. (2023). Detecting hate speech in memes: a review. Artificial Intelligence Review, 56(11), 12833-12851.

[28] Batani, J., Mbunge, E., Muchemwa, B., Gaobotse, G., Gurajena, C., Fashoto, S., ... & Dandajena, K. (2022, April). A review of deep learning models for detecting cyberbullying on social media networks. In Computer Science On-line Conference (pp. 528-550). Cham: Springer International Publishing.