https://africanjournalofbiomedicalresearch.com/index.php/AJBR
Afr. J. Biomed. Res. Vol. 28(2s) (February 2025); 321-328
Research Article

Hate Speech Detection Using Social Media Discourse: A Multilingual Approach With Large Language Model

Muhammad Ahmad[1], Muhammad Usman[1], Sulaiman Khan[1], Muhammad Muzamil[2], Ameer Hamza[2], Muhammad Jalal[2], Ildar Batyrshin[1], Usman Sardar[3*], and Carlos Aguilar-Ibañez[1*]

[1]Centro de Investigación en Computación, Instituto Politécnico Nacional (CIC- PN), Mexico City 07738, Mexico
[2]Department of Artificial Intelligence, Computer Science and Software Engineering, The Islamia University of Bahawalpur, 63100, Pakistan
[3]School of Informatics and Robotics, Institute of Arts and Culture, Lahore 54000, Pakistan

***Corresponding Author:** Usman Sardar[3], Carlos Aguilar-Ibañez[1]
[3]School of Informatics and Robotics, Institute of Arts and Culture, Lahore 54000, Pakistan
[1]Centro de Investigación en Computación, Instituto Politécnico Nacional (CIC- PN), Mexico City 07738, Mexico

Abstract
Online social networks (OSN) and microblogging websites are attracting Internet users and have revolutionized how we communicate with individuals, share their feelings, and exchange ideas across the world with ease. In the extensive age of social media, there is increasing online hate speech, which can provoke violence and contribute to societal division. Hate speech based on race, gender, or religion puts those affected at risk of mental health problems and exacerbates social problems. While current protocols have reduced overt hate speech, subtler forms known as implicit hate speech have emerged, making detection more challenging. This study focuses on hate speech detection using social media discourse, by creating a comprehensive multilingual dataset [25] in Urdu and English and applied multiple machine learning, deep learning, transfer learning, and Large Language model models, such as GPT-3.5 Turbo. By comparing GPT-3.5 Turbo, we identified the effectiveness of large language models in detecting both explicit and implicit forms of hate speech. Our analysis underscores the potential of automated classification systems to reduce reliance on human intervention and to promote constructive online discourse. Our proposed methodology achieved the highest accuracy of 0.91, and achieved the highest performance improvement of 5.81% over transformer models such as BERT. This research adds to the growing body of work on multilingual natural language processing (NLP) and offers insights for reducing hate speech and fostering respectful communication across diverse communities.

Keywords: LLM, GPT, Machine learning, CNN, BERT algorithm, Social media, SVM

Received: 04/01/2025 Accepted: 24/01/2025

DOI: https://doi.org/10.53555/AJBR.v28i2S.6805

1. Introduction

The extensive use of social media has led to increase in online hate speech, which undermines constructive public discourse and, in some cases, may incite violence and extremism [1], [2]. Hate speech, often targeting people based on characteristics such as race, gender, or religion, fosters discrimination and hostility, leading to negative impacts on victims' mental health [3] and exacerbating social tensions and divisions [4]. While stronger regulations have helped limit overt forms of hate speech, they have also led to the emergence of more indirect, subtle expressions of hate, known as "implicit hate speech," which is harder to detect and recognize because of its nuanced nature [5], [6]. Thus, developing effective automated systems to detect both explicit and implicit forms of hate speech is a priority for researchers and society alike [7].

Existing studies have mainly focused on hate speech targeted at specific groups, often based on characteristics like immigration status, gender [8], religion [9] [10], and race [11]. However, as social media platforms have become more popular, they have also raised public awareness of political issues. Supporters of political figures can now access real-time updates on their activities and proposed policies, increasing engagement across the board. Unfortunately, this engagement sometimes prompts people with polarized political views to use social media as a channel for spreading hate speech against those with opposing beliefs. Detecting hate speech is essential to prevent potential violence and discrimination, whether from those spreading it or toward those it targets. This growing trend highlights the need for robust systems to monitor and address hate speech in political discussions, fostering a more respectful online environment.

Recent years have seen the rise of more efficient hate speech detection systems owing to advancements in natural language processing (NLP) and machine learning [20]. With BERT and Roberta, as well as multilingual embedding's, modeling of complex structures of languages and cross-lingual transfer for the training of language models with generalization capability and even for low-resource languages such as Urdu becomes possible[6,7]. Furthermore, the combination of transformer models with some other classifiers, such as SVM and Random Forest, presents avenues for improvement in detection systems in terms of both efficiency and accuracy. [18, 19].

This study makes three significant contributions to the field of hate speech detection. Firstly, we built a dataset related hate speech using a semi supervised learning annotation procedure in Urdu and English. Secondly we employed joint multilingual techniques and combine datasets into a single CSV file, which provide a valuable resource for the future research. Thirdly, we used preprocessing techniques. Fourthly, we utilized advanced feature extraction methods such as TF-IDF, FasText , Glove and Contextual Embedding's using Transformer models to capture both semantic and syntactic features, enhancing the accuracy and context-awareness in hate speech detection and lastly we apply multiple machine learning , deep learning and transformer models [21] to identify the best fit model which gives the higher accuracy in this way we will identify the most effective solution for our hate speech detection task. These techniques, when applied effectively, enable the model to accurately flag harmful content and support efforts to reduce online hate speech. The term multilingual refers to the structure of our dataset, which integrates both English and Urdu data into a single file. The data is organized in an alternating pattern, where two consecutive rows contain English text followed by two rows of Urdu text, and this sequence is repeated throughout the dataset.

This study makes the following Contributions:

1. We applied the schema to develop a comprehensive Hate speech detection dataset in Urdu and English with 8700 samples and applied the joint multilingual techniques as first time.
2. Conduct a comprehensive analysis and performance evaluation using various deep learning and transfer learning techniques and Large Language model, along with comprehensive visualizations.
3. Propose, implement, and evaluate state-of-the-art Large Language model such as GPT-3.5 Turbo designed to automatically detect hate speech in social media discourse, thereby reducing reliance on human intervention through automated classification systems

The remainder of this paper is organized as follows. Section II outlines the literature survey. Section III contains methodology and design. Section IV presents results and analysis. Finally, Section V presents the conclusions of the study.

2. Literature Survey

Yaosheng et al [12] proposed a comprehensive hate speech detection dataset comprising 20,000 entries across nine domains, addressing the lack of resources for Chinese hate speech detection. They introduced a novel Domain-enhanced Prompt Learning (DePL) method to efficiently handle the complexities of domain specificity and data scarcity. Their experimental results indicate that this methodology achieves state-of-the-art performance in both few-shot and full-scale detection scenarios.

Wang et al [13] addresses the challenge of hate speech proliferation in online environments, proposing a method to develop a political hate speech lexicon and train AI classifiers for detection. The authors collected a Chinese hate speech dataset and implemented both deep learning and lexicon-based approaches to enhance detection capabilities. Their framework aims to balance the need for effective hate speech detection while preserving the freedom of speech online.

Alkomah et al [14] discussed a detail and comprehensive review of detection of hate speech in textual the datasets, they extract textual features, and applied machine learning. They identified key themes across 138 relevant studies, noting that many existing approaches do not yield consistent results across different categories of hate speech. The analysis revealed a tendency towards using combined methods, particularly those that integrate multiple deep learning models. Furthermore,

the review highlighted limitations in the available hate speech datasets, noting that many are small and lack reliability for various detection tasks. This study ultimately aims to provide valuable insights and empirical evidence regarding the characteristics of hate speech, assisting the research community in identifying future research directions.

MacAvaney et al [15] discusses the increasing prevalence of hate speech as online content expands, highlighting significant challenges in automated hate speech detection systems. Key difficulties include the nuances of language, varying definitions of hate speech, and the limited availability of robust datasets for training and evaluating detection algorithms. Additionally, many contemporary methods, particularly neural networks, face issues with interpretability, making it challenging for users to understand the reasoning behind their decisions.

Yin et al [16] investigate the challenge of hate speech detection, emphasizing that existing models often fail to generalize to unseen data. The paper summarizes the generalizability issues, identifies reasons for these challenges, and reviews attempts to address them. Finally, it suggests future research directions to enhance the generalization capabilities of hate speech detection systems.

Del et al [17] investigate the spread of harmful campaigns on social networks, focusing on hate speech in comments on public Italian Facebook pages. They create a taxonomy of hate categories and annotate comments, then develop classifiers using Support Vector Machines (SVM) and Long Short Term Memory (LSTM) networks. Their findings demonstrate the effectiveness of these classifiers in recognizing hate speech within the first annotated Italian Hate Speech Corpus.

3. Methodology and Design

3.1 Data collection

The Tweepy API allows us to filter tweets by various criteria, including date, location, language, and tweet ID to facilitate data collection. In this study, Tweepy API were used to collect the tweets, around 50,000 real world tweets related to hate speech samples were collected in both Urdu and English languages which mainly cover the topic such as. The data was gathered from January 2023 to August 2024, focusing on race, ethnicity, religion, gender, or sexual orientation. To ensure the dataset's diversity and comprehensiveness, we collected text samples. After collecting the samples were exactly labeled as either "hate speech" or "not hate speech," in binary classification approach. By developing a balanced and multilingual dataset, the aim was to train robust machine learning [22] [24] such as Support Vector Machine (SVM) Extreme Gradient Boosting (XGB and Logistic Regression (LR) while in deep learning such as convolutional neural network (CNN) with a bidirectional long short-term memory (BiLSTM), transformer models [23] such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Approach (RoBERTa) a and large models capable of effectively identifying and categorizing hate speech in its various forms across multilingual. as seen in figure 1.

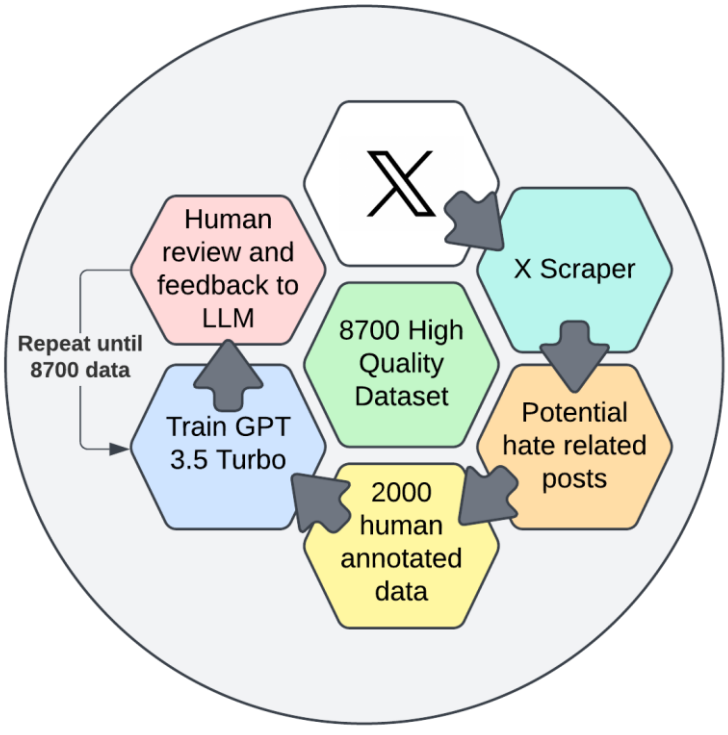

Figure 1. Work flow diagram.

3.2 Annotation.

Accurate and high-quality labeled data is essential for training effective hate speech detection models. To achieve this, we employed a hybrid annotation approach that combines manual labeling with GPT-3.5 Turbo model for efficient and scalable data annotation. We began by manually labeling 2000 instances in binary class to create a foundational dataset, which was then used to train GPT-3.5 Turbo. The trained model was utilized to label additional instances, and through an iterative process of reviewing, correcting, and refining the generated labels, we incrementally expanded the dataset to 8700 accurately labeled instances. This rigorous approach ensured both scalability and quality, leveraging human expertise to enhance GPT-generated annotations and create a reliable resource for hate speech detection.

3.3 Annotation guidelines for initial dataset

3.3.1 Hate speech.

1. If there is racial, ethnic, religious, or gender-based slurs that demean or dehumanize individuals or groups.
2. Language that incites violence, promotes harm, or threatens individuals or groups based on their identity.
3. Terms that reduce individuals or groups to subhuman status or animalistic characteristics, indicating a lack of humanity or empathy.

3.3.2 Not Hate Speech

1. The sentence presents factual information without expressing hostility or bias toward any group.
2. The statement critiques behavior or actions respectfully without targeting individuals or groups based on identity.
3. The sentence engages in dialogue that acknowledges different viewpoints without demeaning any group.

3.4 Pre-Processing

Data preprocessing is an important step that ensures the quality of the dataset for machine learning and deep learning models. Social media data often contain different language so the preprocess start with text cleaning, which involves normalizing the content by removing special characters, links, and unnecessary whitespace, and converting all text to lowercase to achieve uniformity. Next, tokenization is performed to split the text into single words or tokens, allowing for more granular analysis. Given the multilingual nature of the dataset, language-specific tokenization techniques may be applied to account for linguistic variations in Urdu, and English. Additionally, TF-IDF (Term Frequency-Inverse Document Frequency) utilized to reflect the importance of words in the dataset as seen in figure 2. To improve the robustness of the dataset, we employed data augmentation on existing samples. For this purpose we used back translation thereby increasing the dataset's size and diversity. This comprehensive preprocessing approach ensures that the dataset is well-structured and ready for training models, ultimately improving the accuracy and effectiveness of hate speech detection efforts.

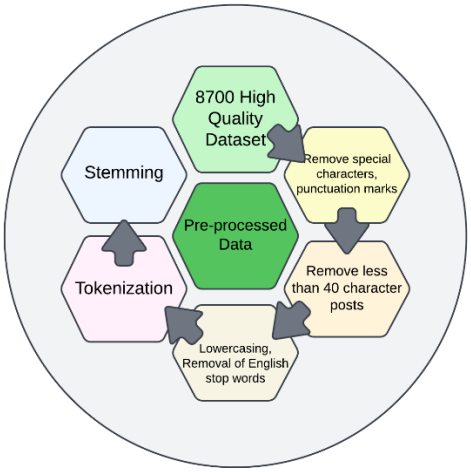

Figure 2. Pre-processing phases

3.5 Application of Models, Training and testing phase

The application of models in this study contains a systematic approach for both training and testing phases to ensure effective performance in hate speech detection. During the training phase, we utilize 80% of data on training and remaining 20% for testing to feed our models, for this purpose we employed various machine learning SVM, XG, and K-NN and deep learning BiLSTM and CNN, transformer such as BERT, Roberta and a large language such as GPT 3.5 Turbo model tailored for text classification. This phase involves hyper-parameter tuning and optimization to enhance model accuracy and generalizability. Once trained, the models undergo a rigorous testing phase, where they are evaluated on a separate validation dataset to assess their performance metrics, including precision, recall, and F1-score.

4. Results and Discussion.

In this section, we will shows the Results of our models based on methodology, alongside a comprehensive discussion of their implications and significance. In experimentations process we train multiple models, including traditional machine learning algorithms and advanced deep learning architectures and large language models, on a diverse dataset of annotated text.

4.1 Results for Machine learning.

Figure 3 shows the performance of machine learning models such as SVM, LR, XGB, and K-NN in a classification task. Each model is assessed based on four key metrics such as precision, recall, F1-score, and accuracy. The SVM shows superior performance with an accuracy of 0.8, signifying a balanced ability to correctly identify positive instances while minimizing false positives. The LR model has slightly lower metrics, with a precision, recall, F1-score, and accuracy all at 0.79. The XGB model matches SVM in precision, recall, and F1-score at 0.8, while also achieving an accuracy of 0.8, indicating its robustness in making accurate

predictions. In contrast, the K-NN model lags significantly behind the others with a precision, recall, F1-score, and accuracy of 0.68, reflecting its limitations in this context. Overall, the results indicate that SVM outperform then all other models.

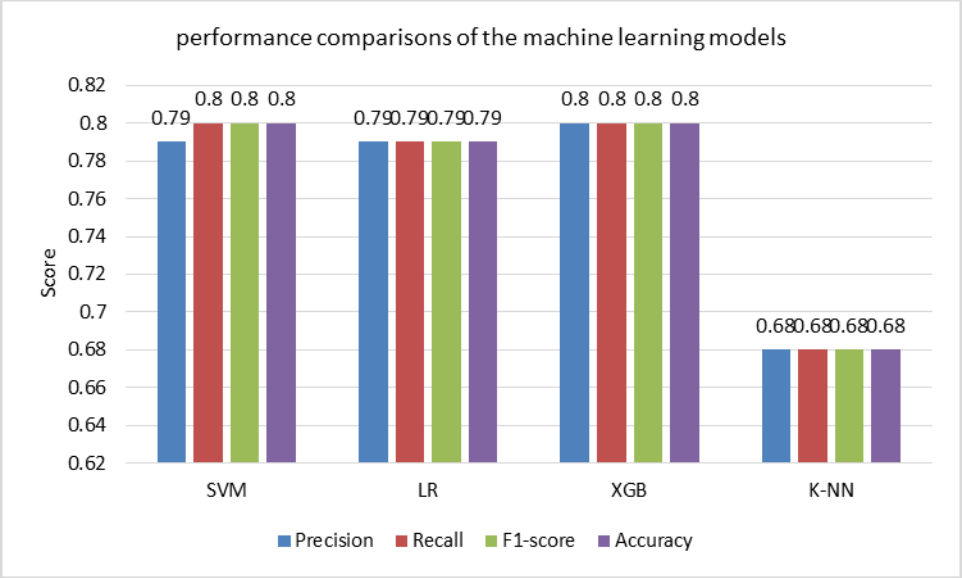

Figure 3. Performance comparisons of the machine learning models.

4.2 Deep learning models.

The Figure 4 illustrate the performance of two deep learning models such as CNN and BiLSTM in hate speech classification task. Both models exhibit strong performance. In comparison, the BiLSTM model performs better across all metrics, with a precision of 0.81, recall of 0.82, F1-score of 0.82, and accuracy of 0.82. This suggests that the BiLSTM not only accurately classifies instances but also captures context better due to its bidirectional nature, leading to improved recall and F1-score.

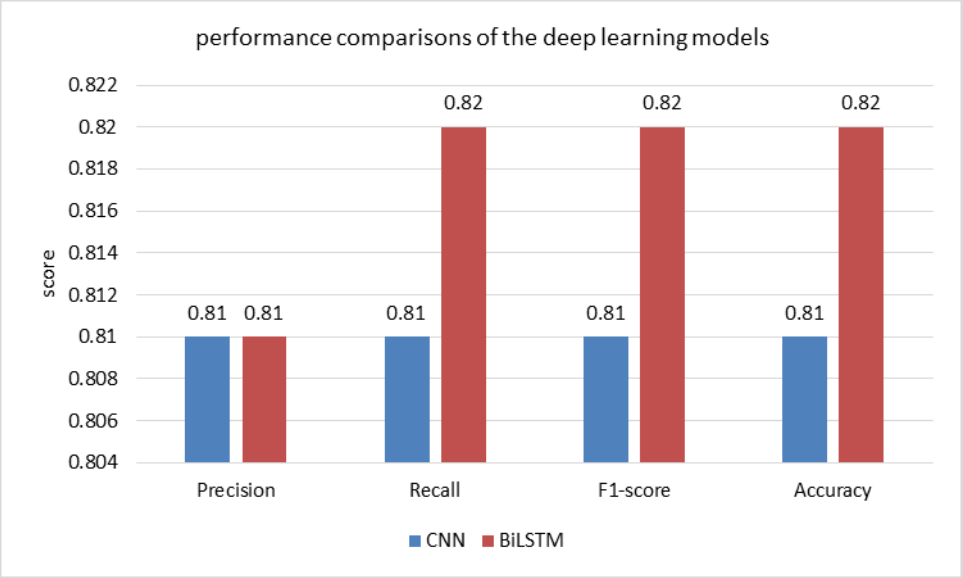

Figure 4.performance comparisons of the deep learning models

4.3 Transformers results

The figure 5 illustrate the performance of two advanced transformer-based models: RoBERTa and BERT. Both models demonstrate strong capabilities, with RoBERTa achieving a precision of 0.85, recall of 0.85, F1-score of 0.84, and accuracy of 0.85. on the other hand BERT, exhibits superior performance across all metrics, with a precision of 0.86, recall of 0.86, F1-score of 0.86, and accuracy of 0.86. This suggests that BERT not only correctly classifies instances with a higher rate but also captures contextual information more effectively, leading to better overall performance in hate speech dataset.

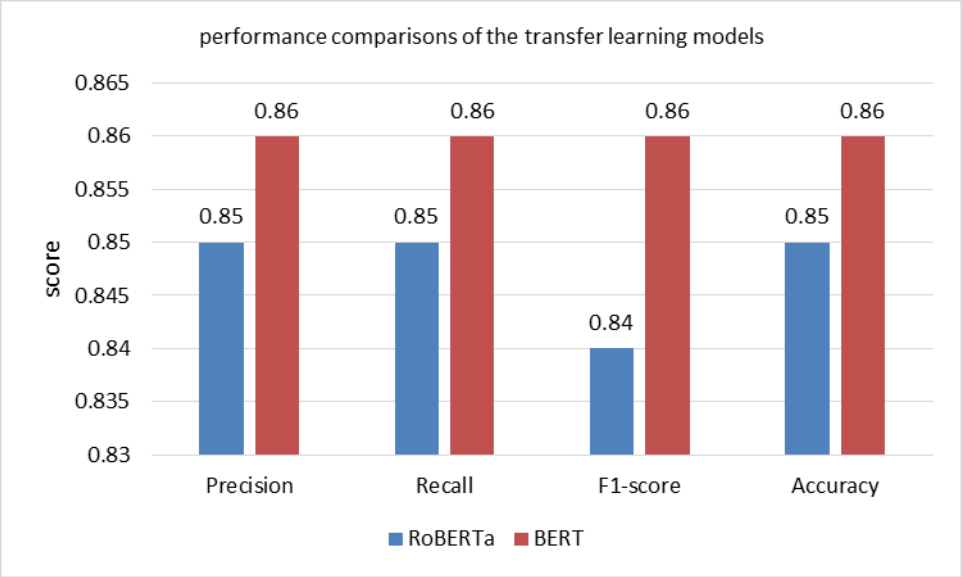

Figure 5.performance comparisons of the Transfer learning models

4.4 Large Language Model

The figure 6 shows the performance of the advance Large Language Model of OpenAi GPT-3.5 Turbo model showcasing high scores with a precision of 0.91, the model proves that 91% of its positive predictions are accurate, indicating a low false positive rate. The recall, also at 0.91, reveals that the model identifies 91% of actual positive instances, showcasing its effectiveness in detecting relevant cases. The F1-score, which combines precision and recall, is also 0.91, indicating a balanced performance between minimizing false positives and maximizing true positives. Finally, an accuracy of 0.91 shows that the model correctly predicts 91% of all instances, reflecting its overall effectiveness in the classification task.

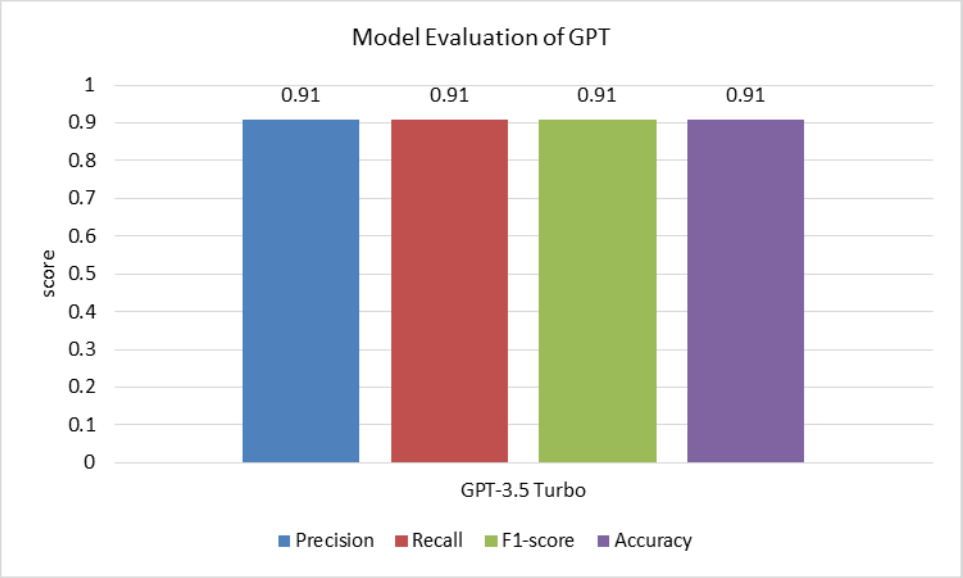

Figure 6.performance comparisons of the Transfer learning models

4.5 Error Analysis

The table 1 display the top-performing models in each learning approaches in a classification task. The SVM) model, achieves a precision of 0.79, a recall of 0.80, an F1-score of 0.80, and an accuracy of 0.80. The BiLSTM model, shows improved performance with a precision of 0.81, recall of 0.82, F1-score of 0.82, and accuracy of 0.82. The BERT model further enhances the results with precision, recall, F1-score, and accuracy all at 0.86. The standout performer is the our proposed model GPT-3.5 Turbo, classified as a which achieves the highest metrics across the board with a precision of 0.91, recall of 0.91, F1-score of 0.91, and accuracy of 0.91 and achieved the highest performance improvement of 5.81% over the Transformer models.

Models	Learning approach	Precision	Recall	F1-score	Accuracy
SVM	Machine leaning	0.79	0.8	0.8	0.8
BiLSTM	Deep learning	0.81	0.82	0.82	0.82
BERT	Transformer	0.86	0.86	0.86	0.86
GPT-3.5 Turbo	LLM	0.91	0.91	0.91	0.91

5. Conclusion

This study demonstrates the effectiveness of using multilingual techniques and large language models for hate speech detection across multiple languages, specifically Urdu and English. By constructing a comprehensive dataset and applying advanced feature extraction and machine learning methods, we have enhanced the accuracy and efficiency of automated systems. Our approach not only contributes valuable resources for future research but also advances the capabilities of hate speech detection, particularly in social media discourse. This work paves the way for more robust, scalable, and context-aware systems that can help mitigate the spread of harmful content online.

Funding: This research did not receive any funding.

Acknowledgements: This work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge support of Microsoft through the Microsoft Latin America PhD.

References

1. N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, ‘‘Hate speech detection with comment embeddings,’’ in Proc. 24th Int. Conf. World Wide Web, May 2015, pp. 29–30.
2. Y. Chen, Y. Zhou, S. Zhu, and H. Xu, ‘‘Detecting offensive language in social media to protect adolescent online safety,’’ in Proc. Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Social Comput., Sep. 2012, pp. 71–80.
3. B. M. Tynes, M. T. Giang, D. R. Williams, and G. N. Thompson, ‘‘Online racial discrimination and psychological adjustment among adolescents,’’ J. Adolescent Health, vol. 43, no. 6, pp. 565–569, Dec. 2008.
4. M. L. Williams and P. Burnap, ‘‘Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data,’’ Brit. J. Criminol., vol. 56, no. 2, pp. 211–238, Mar. 2016.
5. W. Warner and J. Hirschberg, ‘‘Detecting hate speech on the World Wide Web,’’ in Proc. 2nd Workshop Lang. Social Media, 2012, pp. 19–26.
6. H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths, ‘‘A web of hate: Tackling hateful speech in online social spaces,’’ in Proc. Workshop Programme, 2016, p. 1.
7. M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, ‘‘Latent hatred: A benchmark for understanding implicit hate speech,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., 2021, pp. 345–363.
8. V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, et al., "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter", *Proc. 13th Int. Workshop Semantic Eval.*, pp. 54-63, 2019.
9. N. Ousidhoum, Z. Lin, H. Zhang, Y. Song and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis" in arXiv: 1908.11049, 2019.
10. Alfina, R. Mulia, M. I. Fanany and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study", *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, pp. 233-238, Oct. 2017.
11. P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making", *Policy Internet*, vol. 7, no. 2, pp. 223-242, 2015.
12. Yaosheng, Z., Tiegang, Z., Tingjun, Y., & Li, H. (2024). Domain-enhanced Prompt Learning for Chinese Implicit Hate Speech Detection. IEEE Access.
13. Wang, C. C., Day, M. Y., & Wu, C. L. (2022). Political hate speech detection and lexicon building: A study in Taiwan. IEEE Access, 10, 44337-44346.
14. Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. Information, 13(6), 273.
15. MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8), e0221152.
16. Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science, 7, e598.
17. Del Vigna12, F., Cimino23, A., Dell’Orletta, F., Petrocchi, M., & Tesconi, M. (2017, January). Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the first Italian conference on cybersecurity (ITASEC17) (pp. 86-95).
18. Mubarak, H., Darwish, K., & Abdelali, A. (2017). Abusive Language Detection on Arabic Social Media. Proceedings of the First Workshop on Abusive Language Online.
19. Mishra, A., Yannakoudakis, H., & Shutova, E. (2018). Neural Character-based Composition Models for Abusive Language Detection. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018).
20. Ahmad, Muhammad, et al. "Elegante: A Machine Learning-Based Threads Configuration Tool for

SpMV Computations on Shared Memory Architecture." *Information* 15.11 (2024): 685.

21. Ahmad, Muhammad, et al. "Cotton Leaf Disease Detection Using Vision Transformers: A Deep Learning Approach." *crops* 1 (2024): 3.
22. Ahmed, M., Usman, S., Shah, N. A., Ashraf, M. U., Alghamdi, A. M., Bahadded, A. A., & Almarhabi, K. A. (2022). AAQAL: A machine learning-based tool for performance optimization of parallel SPMV computations using block CSR. *Applied Sciences*, *12*(14), 7073.
23. Ullah, F., Ahmed, M., Zamir, M. T., Arif, M., Felipe-Riverón, E., & Gelbukh, A. (2024, March). Optimal Scheduling for the Performance Optimization of SpMV Computation using Machine Learning Techniques. In *2024 7th International Conference on Information and Computer Technologies (ICICT)* (pp. 99-104). IEEE.
24. Ullah, F., Zamir, M., Arif, M., Ahmad, M., Felipe-Riveron, E., & Gelbukh, A. (2024, March). Fida@ DravidianLangTech 2024: A Novel Approach to Hate Speech Detection Using Distilbert-base-multilingual-cased. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages* (pp. 85-90).
25. Ahmad, M., Sardar, U., Humaira, F., Iqra, A., Muzzamil, M., Hmaza, A., ... & Batyrshin, I. (2024). Hope Speech Detection Using Social Media Discourse (Posi-Vox-2024): A Transfer Learning Approach. Journal of Language and Education, 10(4), 31-43.