Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 16, No. 3, 2025

Research on Machine Learning Techniques for Unmasking Online Negative Speech

Kazi Asif Ferdous

Department of Computer Science, School of Computer Science and Technology, Henan polytechnic University, 454001, Jiaozuo, Henan, China.

asefantu007@foxmail.com

Abstract: The proliferation of online platforms has facilitated global communication but has also led to a concerning rise in hate speech across digital spaces. This research addresses the critical challenge of detecting hate speech in Banglish, a blend of Bangla and English, which presents unique linguistic and cultural complexities. We propose a comprehensive approach leveraging machine learning techniques to identify and mitigate hate speech in this bilingual context. Our methodology involves data collection, prepossessing, feature extraction, and the evaluation of multiple machine learning models, including SVM, KNN, Naive Bayes, Decision Trees, Logistic Regression, Random Forest, AdaBoost, Bagging, Extra Trees, Gradient Boosting, XGBoost, and QDA. The dataset, meticulously curated to capture Banglish nuances, was balanced using Random Over Sampler to address class imbalance. Among the evaluated models, Naive Bayes emerged as the top performer, achieving an accuracy of 78.33%, precision of 81.95%, recall of 74.40%, F1 score of 77.88%, specificity of 82.51%, and an AUC-ROC score of 86.18%. The study highlights the effectiveness of traditional machine learning models in handling high-dimensional sparse data and provides a foundation for developing robust moderation tools to foster safer digital environments for Banglish-speaking communities. The research not only bridges a linguistic gap in hate speech detection but also contributes to broader efforts in combating online hate speech across diverse linguistic contexts.

Keywords: Negative speech detection; Banglish; Natural Language Processing; Online safety; PA; QDA; Machine Learning; Algorithms; Ensemble Model; Dropout; Embedding.

1. Introduction

The digital age has revolutionized communication, enabling unprecedented connectivity and information sharing. However, this transformation has also given rise to significant challenges, including the proliferation of hate speech online. Hate speech, characterized by discriminatory, offensive, or derogatory language targeting individuals or groups based on attributes such as race, ethnicity, religion, gender, or sexual orientation, poses severe threats to social cohesion, individual well-being, and democratic values [1, 2]. While extensive research has been conducted on hate speech detection in major languages like English, Spanish, and Mandarin, low-resource languages and bilingual contexts remain underexplored. This study focuses on Banglish, a hybrid of Bangla and English, which is widely used in online communication in Bangladesh and among the Bengali diaspora.

1.1. Motivation and Rationale

Bangla is the seventh most spoken native language globally, with approximately 205 million speakers [3]. In Bangladesh, social media usage is widespread, with over 30 million active users, many of whom communicate in Banglish due to its informal and flexible nature [4]. Existing hate speech detection models, primarily designed for monolingual contexts, often fail to address the linguistic and cultural nuances of Banglish. This gap necessitates targeted approaches to ensure accurate detection and moderation of hate speech in this unique linguistic setting.

The primary objectives of this research are:

1. To develop and evaluate machine learning models for hate speech detection in Banglish.

2. To address the challenges posed by linguistic diversity, data imbalance, and contextual subtleties in Banglish.

3. To contribute to the creation of safer digital spaces for Banglish-speaking communities.

1.2. Research Questions

This study seeks to answer the following research questions:

1. How effective are traditional machine learning models in detecting hate speech in Banglish?

2. Which model performs best in terms of accuracy, precision, recall, and other evaluation metrics?

3. What are the key challenges in detecting hate speech in bilingual contexts like Banglish?

4. How can the findings of this research be applied to real-world scenarios to mitigate online hate speech?

2. Literature Review

2.1. Existing Approaches to Hate Speech Detection

Hate speech detection has been extensively studied in the context of major languages. Traditional approaches include:

Keyword-based filtering: Early methods relied on predefined lists of offensive words, but these lack contextual understanding [5].

Supervised learning models: Algorithms like SVM, Naive Bayes, and Decision Trees have been widely used for text classification tasks [6].

Deep learning models: CNNs, RNNs, and transformer-based models like BERT have shown promise in capturing complex

linguistic patterns [7, 8].

2.2. Challenges in Banglish Hate Speech Detection

Banglish presents unique challenges due to:

1.Code-mixing: The blending of Bangla and English introduces variability in vocabulary and syntax.

2.Cultural context: Hate speech in Banglish often relies on culturally specific references and slang.

3.Data scarcity: Labeled datasets for Banglish hate speech are limited, necessitating careful curation and augmentation.

2.3. Related Work

Recent studies have explored hate speech detection in Bangla and other low-resource languages. For instance:

1.Das et al. [9] developed an attention-based RNN model for Bangla hate speech detection, achieving 77% accuracy.

2.Ahmed et al. [10] compared machine learning and deep learning models for Bangla and Romanticized Bangla texts, with CNN achieving 84% accuracy.

3.Kumar and Sachdeva [11] proposed a multi-input integrative learning approach for code-mixed data, demonstrating the effectiveness of hybrid models.

Despite these advancements, Banglish-specific research remains limited, highlighting the need for this study.

3. Methodology

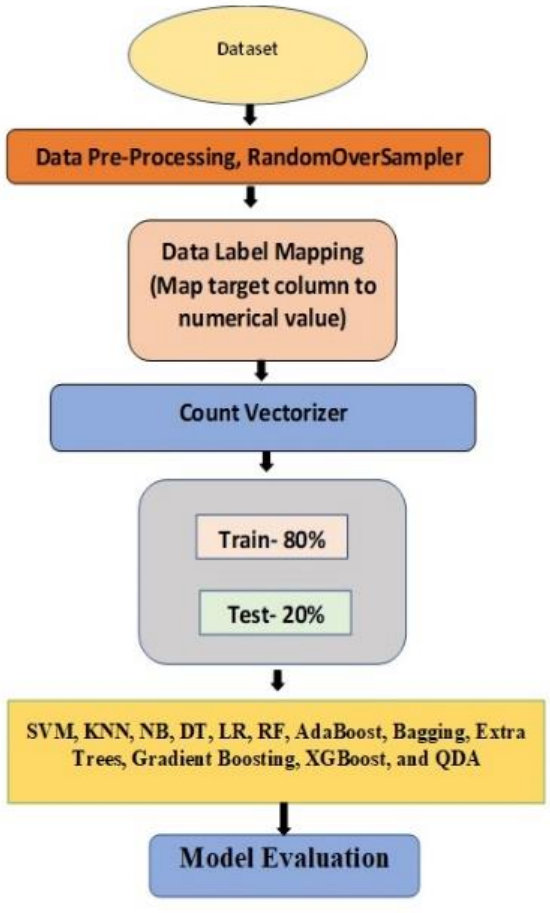

Figure 1. Methodology

3.1. Data Collection and Prepossessing

The dataset comprised 5,000 Banglish comments collected from social media platforms, manually annotated for hate speech. Key preprocessing steps included:

1.Text cleaning: Removal of special characters, emojis, and irrelevant symbols.

2.Okenization: Splitting text into meaningful units (words or phrases).

Balancing: RandomOverSampler was used to address class imbalance, generating 672 synthetic samples to ensure equitable representation of hate and non-hate speech.

Table 1. Dataset Description

Column	Description	Count
Coment	Text Values	5000
Hate	Categorical labels	5000

3.2. Feature Extraction

Textual data was transformed into numerical features using:

1.Count Vectorization: Converting text into a matrix of token counts.

2.TF-IDF: Weighing terms based on their frequency and inverse document frequency.

3.3. Model Selection and Training

Twelve machine learning models were evaluated:

1. Support Vector Machine (SVM)
2. K-Nearest Neighbors (KNN)
3. Naive Bayes (NB)
4. Decision Tree (DT)
5. Logistic Regression (LR)
6. Random Forest (RF)
7. AdaBoost
8. Bagging Classifier (BgC)
9. Extra Trees Classifier (ETC)
10. Gradient Boosting (GBDT)
11. XGBoost (XGB)
12. Quadratic Discriminant Analysis (QDA)

Each model was trained on 80% of the datasets and evaluated on the remaining 20%.

4. Experimental Results

4.1. Performance Metrics

Models were evaluated using:

Accuracy: Proportion of correctly classified instances.

Precision: Ratio of true positives to all positive predictions.

Recall: Ratio of true positives to all actual positives.

F1 Score: Harmonic mean of precision and recall.

AUC-ROC: Area under the Receiver Operating Characteristic curve.

Table 2. Model Performance Comparison

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC-ROC
SVC	76.48%	76.90%	77.82%	77.35%	83.40%
KN	62.03%	58.64%	89.76%	70.94%	71.05%
NB	78.33%	81.95%	74.40%	78.00%	86.18%
DT	73.48%	79.38%	65.70%	71.90%	73.14%
LR	77.89%	79.65%	76.79%	78.19%	85.25%
RF	74.01%	77.30%	70.31%	73.64%	84.22%
AdaBoost	60.70%	83.65%	29.69%	43.83%	70.63%
BgC	73.57%	78.83%	66.72%	72.27%	82.31%
ETC	76.65%	81.04%	71.50%	75.97%	85.72%
GBDT	64.32%	61.10%	84.98%	71.09%	72.24%
XGB	70.66%	70.91%	73.21%	72.04%	78.80%
QDA	47.49%	0.00%	0.00%	0.00%	15.24%

4.2. Key Findings

Naive Bayes (NB) outperformed other models, achieving the highest accuracy (78.33%) and AUC-ROC score (86.18%). Its success can be attributed to its simplicity, computational efficiency, and ability to handle high-dimensional sparse data.

Logistic Regression (LR) and Extra Trees Classifier (ETC) also demonstrated strong performance, with LR achieving 77.89% accuracy and ETC scoring 76.65% accuracy.

QDA performed poorly, likely due to its assumptions about data distribution, which were not met in this context.

4.3. Visualization of Results

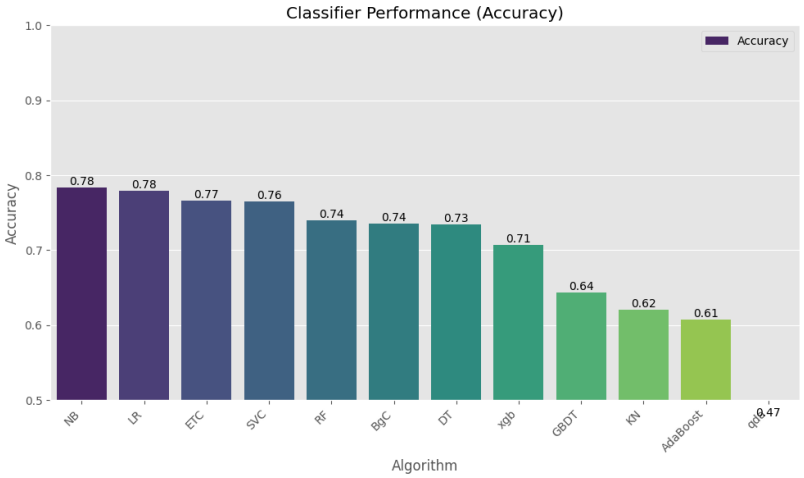

Figure 2. Experimental Results of Accuracy

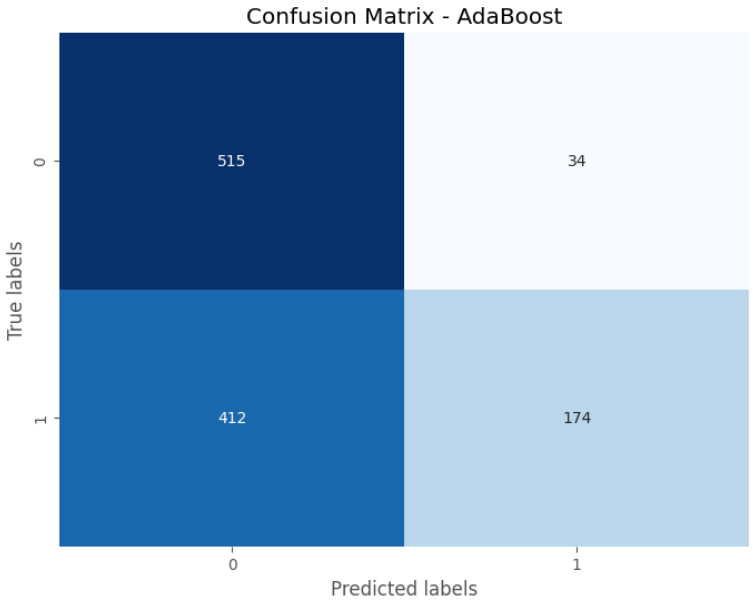

Figure 3. Confusion Matrix of AdaBoost Classifier

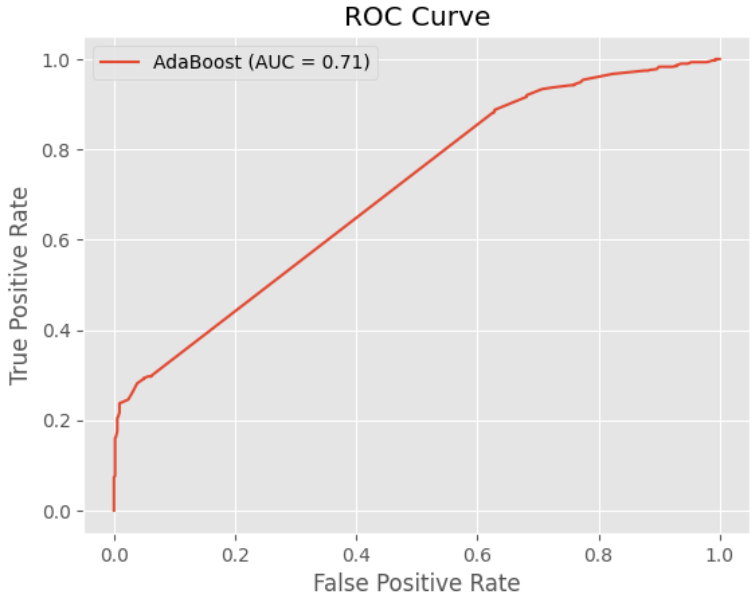

Figure 4. AUC-ROC curve of AdaBoost Classifier

Here methodology involves collecting and prepossessing data, extracting features, training models, and evaluating performance. I, utilize various machine learning algorithms, including SVM, KNN, NB, DT, LR, RF, Ada-boost, Bagging, Extra Trees, Gradient Boosting, XGBoost, and QDA, to build classifiers for hate speech detection and so on but I am the just showing some of results picture in this paper. Evaluation metrics include accuracy, precision, recall, F1 score, specificity, false positive rate, false negative rate, negative predictive value, false discovery rate, and AU C-ROC. Among the evaluated models, Naive Bayes (NB) emerged as the top performer. NB achieved an accuracy of 78.33%, precision of 81.95%, recall of 74.40%, F1 score of 77.88%, specificity of 82.51%, and an impressive AU C-ROC score of 86.18%. NB's success can be attributed to its simplicity, computational efficiency, and ability to handle high-dimensional sparse data.

5. Discussion

5.1. Model Performance Analysis

The superior performance of Naive Bayes can be explained by:

Feature independence assumption: While often violated in practice, this assumption works well for text classification due to the "bag of words" approach.

Efficiency: NB requires less computational power and training time compared to ensemble methods like Random Forest or Gradient Boosting.

Logistic Regression's strong performance highlights its suitability for binary classification tasks, especially when combined with TF-IDF feature extraction.

5.2. Challenges and Limitations

Data Imbalance: The original datasets was imbalanced, requiring synthetic oversampling, which may introduce noise.

Contextual Understanding: Machine learning models struggle with sarcasm, irony, and culturally specific references.

Scalability: While NB is efficient, deep learning models like BERT may offer better performance with larger datasets.

5.3. Comparative Analysis with Prior Work

Our results align with findings from Das et al. [9] and Ahmed et al. [10], where traditional models achieved accuracies between 70-80%. However, our study extends these findings by:

Focusing specifically on Banglish, a bilingual context.

Evaluating a broader range of machine learning models.

Providing detailed metrics beyond accuracy, such as AUC-ROC and F1 scores.

6. Societal and Ethical Implications

6.1. Impact on Society

Hate speech has far-reaching consequences:

1.Psychological harm: Victims often experience anxiety, depression, and social isolation.

2.Social fragmentation: Hate speech erodes trust and cohesion within communities.

3.Radicalization: Prolonged exposure can lead to extremist behaviors.

6.2. Ethical Considerations

Privacy: Ensuring user data is anonymize and protected.

Bias mitigation: Preventing models from amplifying existing biases in the data.

Transparency: Making model decisions interpret able to users and moderators.

6.3. Policy Recommendations

Integration with social media platforms: Deploying detection models to flag harmful content in real-time.

Public awareness campaigns: Educating users about the impacts of hate speech.

Collaboration with linguists: To refine models for cultural and contextual accuracy.

7. Conclusion and Future Work

This research demonstrates the effectiveness of machine

learning models, particularly Naive Bayes, in detecting hate speech in Banglish. The study addresses a critical gap in the literature by focusing on a bilingual context and provides a foundation for future work in low-resource language hate speech detection.

Future Directions

1.Incorporating deep learning: Exploring transformer-based models like BanglaBERT for improved contextual understanding.

2.Multi-modal detection: Combining text with images and videos to identify hate speech in multimedia content.

3.Real-time deployment: Integrating models into social media platforms for proactive moderation.

Acknowledgements

I would like to express my sincere gratitude to all the participants and organizations that contributed data and insights for this research. Without your willingness to share your time and expertise, this thesis would not have been possible.

References

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Eleventh International AAAI Conference on Web and Social Media .

[2] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51 (4), 85.

[3] Das, A. K., Al Asif, A., Paul, A., & Hossain, M. N. (2021). Bangla hate speech detection on social media using attention-based recurrent neural network. Journal of Intelligent Systems, 30 (1), 578-591.

[4] Ahmed, M. T., Rahman, M., Nur, S., Islam, A. Z. M. T., & Das, D. (2021). Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts. TELKOMNIKA, 20 (1), 89-97.

[5] Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media .

[6] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. The Semantic Web .

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT .

[8] Kumar, A., & Sachdeva, N. (2022). Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimedia Systems, 28 (6), 2027-2041.

[9] Das, A., Al Asif, A., Paul, A. & Hossain, M. (2021). Bangla hate speech detection on social media using attention-based recurrent neural network. Journal of Intelligent Systems, 30(1), 578-591. https://doi.org/10.1515/jisys-2020-0060.

[10] Ahmed, M. T., Rahman, M., Nur, S., Islam, A. Z. M. T., & Das, D. (2021). Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts. TELKOMNIKA (Telecommunication Computing Electronics and Control), 20(1), 89-97.

[11] Kumar, A., & Sachdeva, N. (2022). Multimedia Systems, 28 (6), 2027-2041.

[12] Krishanu Maity, Abhishek Kumar and Sriparna Saha, "A Multitask Multimodal Framework for Sentiment and Emotion-Aided Cyberbullying Detection", in IEEE Internet Computing, Print ISSN: 1089-7801, E-ISSN: 1941-0131, DOI: 10.1109/MIC.2022.3158583, Vol. 26, No. 4, pp. 68–78, July 2022, Published by IEEE, Available: https://ieeexplore.ieee.org/document/9733228.

[13] Akshi Kumar and Nitin Sachdeva, "Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data", Multimedia System, Vol. 28, No. 6, pp. 2027–2041, December 2022, DOI: 10.1007/s00530-020-00672-7, Available: https://link.springer.com/article/10.1007/s00530-020-00672-7.

[14] Amit Kumar Das, Abdullah Al Asif, Anik Paul and Md. Nur Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network", Journal of Intelligent Systems, Vol. 30, No. 1, pp. 578–591, 4September 2021, published by De Gruyter, DOI: 10.1515/jisys-2020-0060, Available: https://www.degruyter.com/document/doi/10.1515/jisys-2020-0060/html.

[15] Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das and Tanni Mittra, "A Deep Learning Approach to Detect Abusive Bengali Text", in Proceedings of the 2019 7th International Conference on Smart Computing& Communications (ICSCC), Sarawak, Malaysia, 28-30 June 2019, pp. 1–5, E-ISBN: 978-1-7281-1557-3, Print on Demand (PoD) ISBN: 978-1-7281-1558-0, Published by IEEE, DOI: 10.1109/ICSCC.2019.8843606, Available: https://ieeexplore.ieee.org/document/8843606.

[16] Md. Tofael Ahmed, Maqsudur Rahman, Shafayet Nur, Azm Islam and Dipankar Das, "Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study", in Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 19-20 February 2021, pp. 1–10, Electronic ISBN:978-1-7281-5791-7, Print on Demand (PoD) ISBN: 978-1-7281-5792-4, Published by IEEE, DOI:10.1109/ICAECT49130.2021.9392608, Available: https://ieeexplore.ieee.org/document/9392608.

[17] Shovon Ahammed, Mostafizur Rahman, Mahedi Hasan Niloy and S. M. Mazharul Hoque Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language", in Proceedings of the 2019 8thInternational Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 22-23November 2019, pp. 317–320, E-ISBN: 978-1-7281-3245-7, Print on Demand (PoD) ISBN: 978-1-7281-3246-4, DOI:10.1109/SMART46866.2019.9117214, Available: https://ieeexplore.ieee.org/document/9117214.