Temporal Analysis and Sentiment-Aware Detection of Hate Speech and Cyberbullying

Het Shah
Department of Information & Communication Technology
Pandit Deendayal Energy University
Ahmedabad, India
het.sict21@sot.pdpu.ac.in

Mihir Kosrekar
Department of Information & Communication Technology
Pandit Deendayal Energy University
Ahmedabad, India
mihir.kict21@sot.pdpu.ac.in

Yashrajsinh Dodiya
Department of Information & Communication Technology
Pandit Deendayal Energy University
Ahmedabad, India
yashraj.dict21@sot.pdpu.ac.in

***Abstract*—Hate speech and cyberbullying have become prevalent issues on social media platforms, posing significant risks to mental health, safety, and social harmony. This paper introduces a comprehensive approach for detecting hate speech and cyberbullying by combining sentiment and temporal analysis with advanced machine learning and deep learning models, specifically BERT. Our method addresses the evolving and context-dependent nature of online language by leveraging BERT's contextual embeddings alongside traditional machine learning models. This system aims to improve detection accuracy while maintaining a balance between freedom of expression and content moderation. We also explore the impact of temporal patterns on hate speech and analyze sentiment to enhance detection robustness. Our proposed solution achieves high performance across various metrics, indicating its effectiveness in identifying harmful online content.**

***Index Terms*—Hate Speech Detection, Cyberbullying, Sentiment Analysis, Temporal Analysis, Machine Learning, BERT.**

I. Introduction

The proliferation of social media has facilitated communication but has also enabled the rise of harmful behaviors, including hate speech and cyberbullying. These issues negatively affect individuals and communities, often leading to psychological harm and perpetuating social divisions. Existing detection models struggle with accurately identifying hate speech due to the complex and evolving nature of online language, which includes sarcasm, mixed languages, and context-dependent meanings.

This study aims to address these challenges by building a robust and adaptive system capable of detecting hate speech and cyberbullying with high accuracy. Our approach integrates sentiment analysis, which aids in understanding the emotional tone of the text, and temporal analysis, which identifies patterns in hate speech occurrence over time. By leveraging these insights, we develop a detection system that is both effective and adaptable. We implement a suite of models, including BERT, which is known for its powerful contextual embeddings, and traditional machine learning classifiers, to explore their performance across various scenarios.

II. Problem Statement

Existing hate speech detection systems often fail to capture the nuanced and context-dependent nature of online language, particularly with the inclusion of sarcasm, ambiguous expressions, and mixed languages. Furthermore, the temporal dynamics of hate speech, where certain events or times may correlate with spikes in hate speech, are often overlooked. This project addresses the following core challenges:

- **Language Complexity and Evolution**: Hate speech manifests in various forms, including slang, sarcasm, and multilingual phrases, making detection challenging for static models.
- **Contextual Understanding**: Traditional models lack the contextual depth to accurately interpret the intent behind messages, especially when hate speech is veiled.
- **Temporal and Sentiment Insights**: Limited integration of temporal analysis and sentiment polarity reduces the robustness of current models in detecting spikes and patterns in hate speech.

Our goal is to develop an advanced detection system that addresses these challenges by integrating BERT, sentiment analysis, and temporal insights, achieving high detection accuracy without compromising the subtlety of language understanding.

III. Literature Review

Table I provides a summary of previous research on hate speech and cyberbullying detection methods, showcasing models like CNN, LSTM, BERT, and other hybrid approaches along with their key findings and limitations.

IV. Methodology

Our methodology consists of data collection, preprocessing, exploratory data analysis (EDA), model development, and evaluation. The dataset is sourced from various platforms, including Twitter, Reddit, and YouTube, containing labeled hate speech and non-hate speech content.

TABLE I
SUMMARY OF LITERATURE ON HATE SPEECH DETECTION TECHNIQUES

Title	Authors	Year	Techniques Used	Key Findings and Limitations
Cyberbullying Detection in Social Networks Using Deep Learning-Based Models	M. Dadvar, K. Ecker [1]	2018	CNN, LSTM	Achieved 85% accuracy; limited data variety affects generalization.
Transfer Learning for Hate Speech Detection in Social Media	L. Yuan, T. Wang [2]	2023	BERT, Transfer Learning	Improved detection accuracy to 92% on target datasets; dependency on pre-trained models.
Sentiment Analysis for Cyberbullying Detection in Twitter	C. P. Theng, N. F. Othman [3]	2021	Naive Bayes	Sentiment analysis achieved an F1 score of 0.74; limited performance on nuanced cases.
Detecting Hate Speech using BERT and Hate Speech Word Embedding	H. Saleh, A. Alhothali [4]	2023	BERT, Word2Vec	BERT model reached 89% accuracy; requires high computational power.
A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter	A. Muneer, S. M. Fati [5]	2020	SVM, Random Forest	SVM performed best with 83% accuracy; limited ability to handle contextual nuances.
Exploring Data Augmentation for Gender-Based Hate Speech Detection	M. A. Ibrahim, S. Arifin, E. S. Purwanto [6]	2023	Data Augmentation, Random Forest	Increased detection accuracy from 68% to 75%; generalizability concerns.
Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods	A. Toktarova, D. Syrlybay [7]	2023	Random Forest, CNN, LSTM	Hybrid method achieved 90% accuracy; high computational costs.
AI-Enabled Cyberbullying-Free Social Networks in Smart Cities	A. Al-Marghilani [8]	2022	Reinforcement Learning, Sentiment Analysis	Created 90% cyberbullying-free environments in simulations; implementation challenges in real-world scenarios.
BullyNet: Unmasking Cyberbullies on Social Networks	A. S. Srinath, H. Johnson [9]	2021	Graph Convolutional Networks, Logistic Regression	GCN model achieved 88% accuracy; limited scalability for subtle cases.
Towards a Cyberbullying Detection Approach using Contrastive Self-supervised Learning	L. M. Al-Harigy, H. A. Al-Nuaim [10]	2024	Contrastive Self-supervised Learning, Data Augmentation	Reached 84% accuracy with improved generalization; requires significant training.

A. Data Collection

The data collection process involved gathering hate speech-related comments and posts from online platforms. The dataset contains both English and Hinglish (a hybrid of Hindi and English) text samples, categorized as hate speech or non-hate speech.

B. Preprocessing

To prepare the dataset, several preprocessing steps were performed:

- **Tokenization**: Breaking text into individual tokens (words or subwords).
- **Cleaning**: Removing noise such as URLs, special characters, and stop words.
- **Labeling**: Assigning labels for hate and non-hate speech based on predefined criteria.

C. Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns in the data. Key analyses included:

- **Hate Speech vs Non-Hate Speech Distribution**: Analysis of class distribution using pie charts.
- **Temporal Distribution**: Analyzing monthly and hourly patterns in hate speech occurrence.
- **Sentiment Distribution**: Sentiment analysis to determine the polarity of hate speech versus non-hate speech.

V. MODEL DEVELOPMENT

This project implemented a range of models for hate speech detection, including traditional machine learning models and BERT for deep learning.

A. Machine Learning Models

To establish a performance baseline, we implemented several traditional machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and Decision Tree classifiers. Each model was trained on a balanced dataset containing hate speech and non-hate speech samples. The models were evaluated using accuracy, F1 score, and AUC-ROC metrics to assess their effectiveness.

The Support Vector Machine (SVM) classifier achieved an accuracy of 88.5% and an F1 score of 0.87, showcasing strong

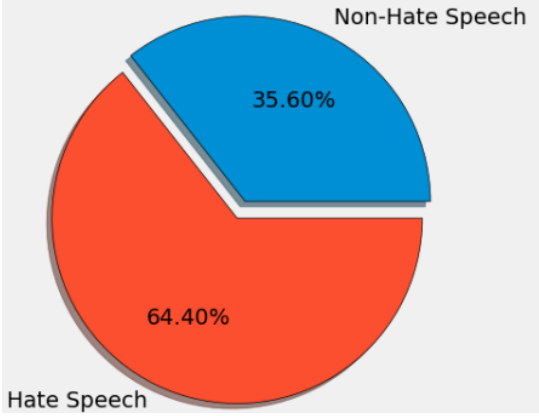

Fig. 1. Distribution of Hate Speech and Non-Hate Speech in Dataset

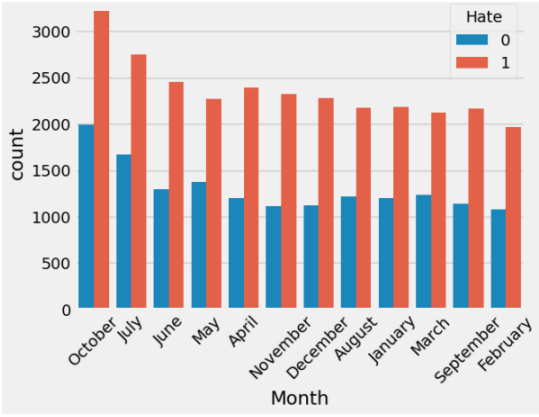

Fig. 2. Monthly Distribution of Hate Speech and Non-Hate Speech

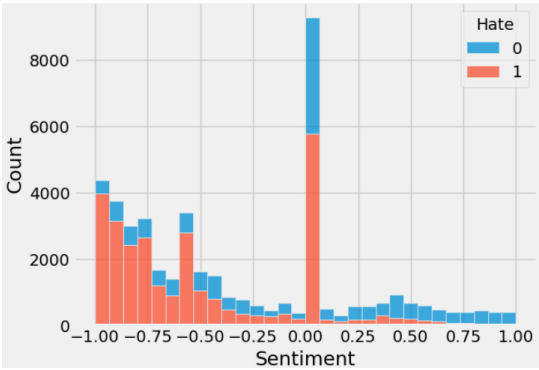

Fig. 3. Sentiment Analysis of Hate Speech and Non-Hate Speech

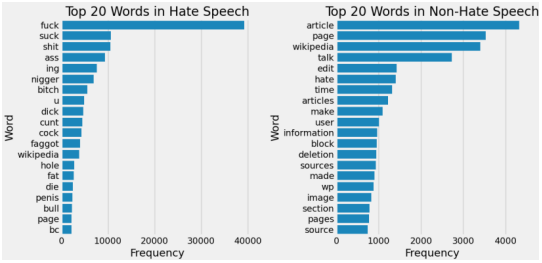

Fig. 4. Most Frequent Words in Hate Speech and Non-Hate Speech

Fig. 5. Word Clouds for Hate Speech and Non-Hate Speech

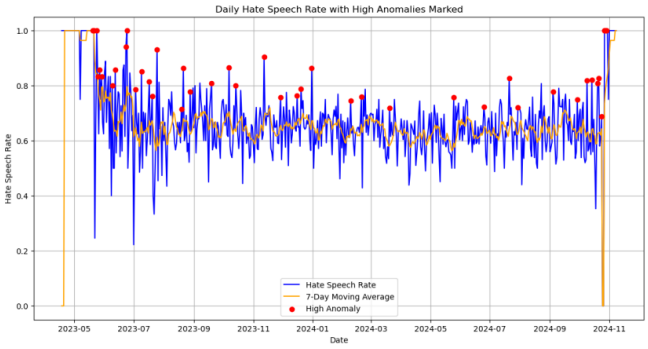

Fig. 6. Anomalies in Hate Speech Detection

performance in binary classification tasks. Random Forest, known for its robustness in handling high-dimensional data, performed comparably with an accuracy of 86.8% and an F1 score of 0.86. Logistic Regression, while slightly less effective in handling complex patterns, achieved an accuracy of 84.5% and an F1 score of 0.84. Decision Tree, due to its susceptibility to overfitting, had a lower accuracy of 80.3% but demonstrated interpretability, which is valuable for understanding feature importance.

Despite these results, traditional machine learning models showed limitations in accurately capturing the nuanced context and semantics in hate speech, necessitating the adoption of a more contextually aware model, such as BERT, to improve overall detection performance.

B. BERT Model

The Bidirectional Encoder Representations from Transformers (BERT) model was fine-tuned for the hate speech detection task due to its capacity for capturing contextual nuances in text. BERT's architecture, which includes bidirectional transformers, enables the model to consider the entire context of a word within a sentence, thus improving its ability to handle complex language structures like sarcasm, mixed languages, and sentiment polarity.

We utilized the 'bert-base-uncased' variant, fine-tuning it on the labeled hate speech dataset. The fine-tuning process involved feeding BERT tokenized text samples and training it with cross-entropy loss over several epochs. The BERT model demonstrated superior performance, achieving an accuracy of 98.91%, an F1 score of 0.9891, and an AUC-ROC of 1.0. This high level of accuracy illustrates BERT's effectiveness in interpreting complex language patterns and detecting nuanced expressions of hate speech. Moreover, BERT's embeddings enhanced our model's sensitivity to sentiment and context,

proving crucial for distinguishing between hateful and non-hateful messages in varied settings.

VI. Results and Analysis

A. Model Performance

Table II shows model performance in terms of accuracy, F1 score, and AUC-ROC. The BERT model outperformed traditional models, achieving high precision and recall.

TABLE II
Performance Metrics for Different Models

Model	Accuracy	F1 Score	AUC-ROC
BERT	98.91%	0.9891	1.00
Decision Tree	95.76%	0.9670	0.96
Random Forest	95.65%	0.9660	0.99

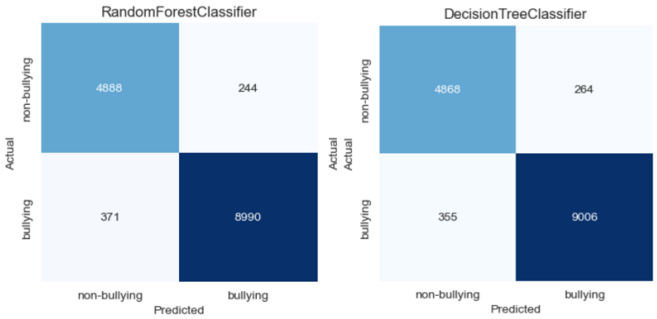

Fig. 7. Confusion Matrix

VII. Conclusion

This study presents a multi-faceted approach to hate speech and cyberbullying detection by incorporating machine learning and deep learning models, specifically BERT, along with sentiment and temporal analyses. By addressing the unique challenges of online language—such as its evolving, context-dependent nature—our system achieved high detection accuracy, particularly with the BERT model, which excelled in identifying subtle forms of hate speech. The system's integration of temporal and sentiment analysis further enhances its ability to adapt to fluctuating patterns in hate speech.

Future work could involve deploying this detection system in real-time applications, enabling platforms to monitor and moderate content effectively. Additionally, exploring multilingual capabilities and adapting the model to cultural variations in language could expand its utility across diverse online communities, contributing to a safer and more inclusive digital environment.

References

[1] M. Dadvar and K. Ecker, "Cyberbullying Detection in Social Networks Using Deep Learning-Based Models," 2018.
[2] L. Yuan and T. Wang, "Transfer Learning for Hate Speech Detection in Social Media," 2023.
[3] C. P. Theng and N. F. Othman, "Cyberbullying Detection in Twitter Using Sentiment Analysis," 2021.
[4] H. Saleh and A. Alhothali, "Detection of Hate Speech using BERT," 2023.
[5] A. Muneer and S. M. Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," 2020.
[6] M. A. Ibrahim, S. Arifin, and E. S. Purwanto, "Exploring Data Augmentation for Gender-Based Hate Speech Detection," 2023.
[7] A. Toktarova and D. Syrlybay, "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," 2023.
[8] A. Al-Marghilani, "AI-Enabled Cyberbullying-Free Social Networks in Smart Cities," 2022.
[9] A. S. Srinath and H. Johnson, "BullyNet: Unmasking Cyberbullies on Social Networks," 2021.
[10] L. M. Al-Harigy and H. A. Al-Nuaim, "Towards a Cyberbullying Detection Approach using Contrastive Self-supervised Learning," 2024.