Department Of Computer Science and Engineering

BHARATI VIDYAPEETH'S COLLEGE OF
ENGINEERING
PASCHIM VIHAR, NEW DELHI

MINI PROJECT SYNOPSIS

For the partial fulfilment of the degree of Bachelor of Technology in
Computer Science and
Engineering
(Session 2021-2025)

Under the supervision of:
Mrs. Rachna Narula
Assistant Professor
Bharati Vidyapeeth College of
Engineering

Submitted by: (CSE-2)
Yash Rakesh(20511502721)
Kashvi Malik(20611502721)
Tanish Gupta(20811502721)
Daksh Balyan (20911502721)

Introduction

In recent years, the proliferation of social media platforms has led to an unprecedented surge in online communication. While these platforms have enabled the exchange of ideas and information on a global scale, they have also witnessed the alarming rise of hate speech, a phenomenon characterized by the use of discriminatory, offensive, or threatening language targeted at individuals or groups based on their race, religion, ethnicity, gender, or other protected characteristics.

1. What is Hate Speech

Hate speech covers a broad spectrum of expressions that belittle, degrade, or encourage violence against a specific person or group. It goes against the core values of inclusiveness, diversity, and respecting human rights that are vital in a democratic society. What makes hate speech dangerous is its ability to worsen social conflicts, encourage hostility, and, in severe instances, even provoke violent actions.

2) Why is it Important

Detecting hate speech in textual data from social media platforms in Hindi or any language is crucial for several reasons:

- Promoting Online Safety: Hate speech can create a hostile online environment that discourages people from expressing their opinions freely. Detecting and mitigating hate speech can help foster a safer online space for all users.
- Protecting Vulnerable Communities: Hate speech often targets specific racial, ethnic, religious, or other marginalized groups. Detecting hate speech can help protect these vulnerable communities from harm and discrimination.
- Preventing Real-world Harm: Hate speech can escalate from online platforms to real-world violence and discrimination. Identifying and addressing hate speech early can help prevent such incidents.
- Maintaining User Trust: Users expect social media platforms to provide a safe and respectful environment. Detecting and addressing hate speech helps maintain user trust and encourages more people to use these platforms.
- Reducing Toxicity: Hate speech contributes to the toxicity of online discussions. By identifying and addressing hate speech, platforms can reduce toxicity and promote more constructive conversations.

3. Challenges in Hate Speech Detection in Hindi Language

While significant strides have been made in hate speech detection in languages like English, the same cannot be said for Hindi, one of the most widely spoken languages globally. This research endeavor confronts several unique challenges:

a. Limited Data Availability

Compared to languages like English, resources for Hindi language processing are relatively scarce. This scarcity of annotated datasets poses a significant obstacle in training and fine-tuning machine learning models for hate speech detection in Hindi.

b. Linguistic Complexity

Hindi is a morphologically rich language with a complex script. The presence of compound words, idiomatic expressions, and regional variations adds an extra layer of complexity to the task of hate speech detection.

c. Contextual Ambiguity

Hindi, like many other languages, often relies on contextual cues for accurate interpretation. Detecting hate speech necessitates a nuanced understanding of cultural, social, and historical contexts, which may not always be straightforward for automated systems.

d. Code-Mixing

Social media interactions in Hindi often involve code-mixing, wherein multiple languages, including English, are used within the same sentence or phrase. This further complicates the task of hate speech detection, as models must be adept at handling multilingual content.

In light of these challenges, this research endeavors to bridge the gap in hate speech detection by focusing specifically on the Hindi language. By addressing these unique hurdles, we aim to contribute to a safer, more inclusive online environment for Hindi-speaking communities.

Problem Statement

Hate Speech detection of textual data from social media platforms in Hindi language.

- The prevalence of hate speech on social media platforms has become a major concern globally, contributing to the rise of online toxicity, cyberbullying, and social discord.

- While extensive research and development have been conducted in the domain of hate speech detection, a noticeable gap exists in the specific context of the Hindi language. The majority of existing studies and tools primarily focus on English, leaving a significant void in the effective identification and mitigation of hate speech in Hindi, which is one of the most widely spoken languages in the world.

Literature Review

SNO	TITLE	METHODOLOGY	LIMITATION
1)	TABHATE: A Target-based Hate Speech Detection Dataset in Hindi	Used lexicon-based approach and XLM-RoBERTa for tweet classification.	Limited to Hindi, may not apply to other languages, cultural/regional variations not considered.
2)	Challenges for Hate Speech Recognition System: Approach based on Solution	Pre-processing: Case folding, tokenization, and punctuation removal. Feature identification: Unigram to 5-gram TF-IDF counts. Multi-view SVM model: Utilizing different feature types.	Evolving attitudes and context. Closed loop systems aiding evasion. Ethical concerns with user data. Future hate speech identification complexity.
3)	Machine Learning based Automatic Hate SpeechRecognition System	Data collection: Public hate speech tweets gathered. Text pre-processing: Irrelevant information removed. Feature engineering: Text converted to numerical vectors using software. Data splitting: Dataset divided into training and test sets.	Limited dataset: Relied on a specific hate speech dataset. Language dependency: Focused on English hate speech detection. Generalizability: Findings may not apply to all platforms. Performance evaluation: Metrics and classifiers were specific.
4)	HateCheckHIn: Evaluating Hindi Hate Speech Detection Model	Data Collection: Obtained data from reliable sources. Preprocessing: Cleaned and formatted data. Feature Extraction: Extracted key features. Model Selection: Choose suitable models for analysis.	Limited Data Availability: Restricted by data availability. Time Constraints: Limited time for data work. Scope: Focused on specific variables.
5)	Hate and Offensive Speech Detection in Hindi Twitter Corpus	Feature extraction: Employed TF-IDF, Bag of Words, and Word2Vec techniques. Classification models: Utilized Logistic Regression, Naïve Bayes, SVM, and Random Forest.	Focus on Hindi Twitter hate speech detection. Limited dataset domain and language. Model performance may vary with different datasets.

		Dataset split: Partitioned into 80% for training and 20% for testing.	
6)	Social Network detection system for amharic language	Literature review for relevant studies. Research plan and data collection methods. Data collection via structured questionnaires. Data analysis using statistical software.	Limited sample size impacts generalizability. Time constraints on data collection. Self-report data bias due to social desirability. Lack of control over external factors.
7)	A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere	Collected labeled tweets (hateful, abusive, normal). Compared CNN, GRU, CNN + GRU model performance. Evaluated BERT for Arabic hate speech detection. Applied data preprocessing.	Arabic hate speech dataset may not generalize. Misclassification of non-offensive tweets. Challenge in classifying context-limited tweets. Performance varies with different datasets.
8)	Deep Learning for Hate Speech Detection in Tweets	Experimented with various deep learning architectures. Employed a 16K annotated tweet benchmark dataset. Evaluated using 10-fold cross-validation. Calculated precision, recall, and F1-scores for comparison.	Focused solely on Twitter hate speech. Dataset may not encompass all hate speech types. Method performance variability on diverse datasets. Word embeddings may not capture all hate speech nuances.
9)	A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection	Data Collection: Gathered information from reliable sources. Literature Review: Examined existing studies. Research Design: Created a structured plan. Data Analysis: Employed statistical methods for analysis.	Sample size limited. Time constraints on data work. Findings may not generalize.
10)	ABMM: Arabic BERT-Mini Model for Hate-Speech	Data collection: Twitter tweets gathered with specific hashtags.	Unbalanced dataset: Challenges in equal representation of hate and non-hate tweets.

	Detection on Social Media	Pre-processing: Removed punctuation, normalized Arabic text, eliminated usernames, URLs, and hashtags. Model development: Created Arabic BERT-Mini Model (ABMM) for hate speech detection. Evaluation: Compared ABMM with traditional ML models and state-of-the-art approaches.	Lack of additional features: Ignored factors like emoji descriptions. Interpretability: ABMM lacks explanations for classification decisions. Hardware constraints: Limited by memory and CPU for deep learning layers.
11)	arHateDetector: detection of hate speech from standard and dialectal Arabic Tweets	Data Collection: Gathered relevant data. Preprocessing: Cleaned and standardized data. Model Selection: Choose appropriate models. Model Training: Trained models using preprocessed data.	Limited Data Availability: Constrained by data availability. Time Constraints: Research conducted within a specific timeframe. Resource Limitations: Limited computational resources.
12)	Transformers and Ensemble methods: A solution for Hate Speech Detection in Arabic languages.	Six transformer models for hate speech detection. Two ensemble methods to combine models. Cross-validation for performance evaluation. Best model selection based on F1-score.	Limited computational resources for model size. Dataset focus on COVID-19 disinformation restricts generalizability. Models trained on Arabic text limit language applicability. Evaluation metric focuses solely on F1-score for the positive class.
13)	AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset	Data Collection: Gathered tweets using specific keywords. Data Annotation: Expert-labeled tweets. Preprocessing: Basic text preprocessing applied. Model Training: Pretrained transformer models used for classification.	Single expert annotator per tweet. Small dataset (10,828 tweets) may impact model performance. Focused on Arabic tweets, limiting generalizability.
14)	Hate Speech Detection in Hindi language using BERT and Convolution Neural Network	Preprocess text data: Remove URLs, usernames, punctuation. Translate emoticons to Hindi descriptions. Use pretrained BERT encoder for contextualized embeddings.	Limited Hindi hate speech datasets. Imbalanced distribution of hate and non-hate classes. Addressed imbalance with oversampling. Evaluation based on f1-score due to dataset imbalance.

		Implement CNN with parallel convolution filters.	
15)	QutNocturnal@HASOC'19: CNN for Hate Speech and Offensive Content Identification in Hindi Language	Data Collection: Gathered Hindi tweets. Preprocessing: De-identified and removed URLs. Word Embedding: Used Word2Vec for word vectors. Model Architecture: Created a custom CNN model.	Limited training data: 4665 labeled tweets. Noise in tweets: Misspellings and noise present. Language-specific model: Designed for Hindi.

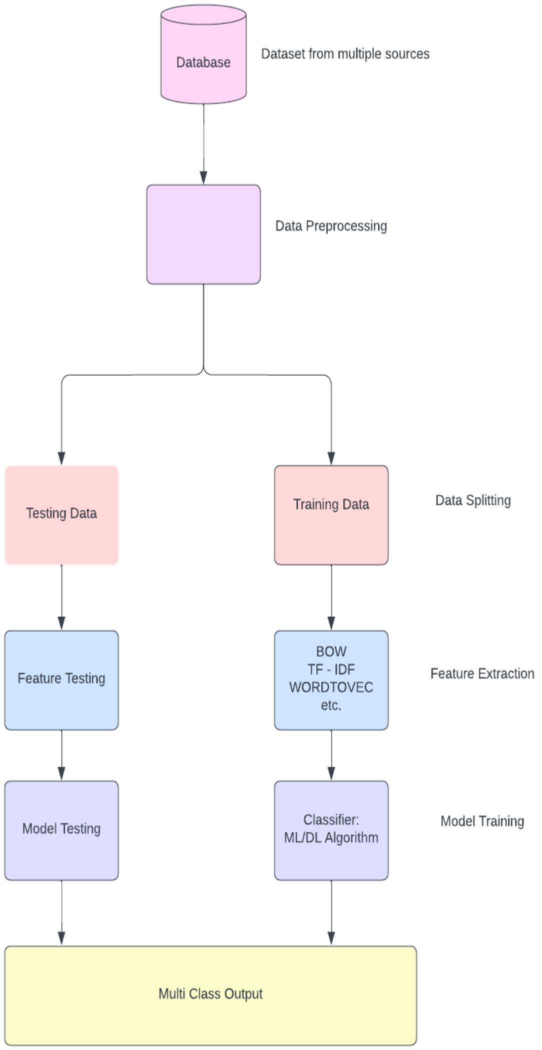
Objectives

1. To design a meticulously curated and annotated dataset with a focus on target-based hate speech detection, particularly within the context of the Hindi language.
   Develop a Robust Hate Speech Detection Algorithm for Hindi Text

2. Implement Advanced NLP Techniques for Hindi
   Apply advanced NLP techniques, including sentiment analysis and contextual understanding, specifically tailored to the complexities of the Hindi language.

3. Evaluate Model Performance and Fine-tune as Necessary
   Rigorously assess the hate speech detection model's accuracy, sensitivity, and specificity using a variety of evaluation metrics. Make necessary adjustments to improve its effectiveness.

Proposed Methodology/ Workflow Diagram



References

[1] M. Das, P. Saha, B. Mathew, and A. Mukherjee, "HateCheckHIn: Evaluating Hindi Hate Speech Detection Models," in Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, 2022, pp. 5378-5387.

[2] J. Mishra, M. Vidgen, D. Nguyen, R. Waseem, H. Margetts, and J. Pierrehumbert, "Challenges and Frontiers in Abusive Content Detection," in Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 2019, pp. 80-93.

[3] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 2016, pp. 88-93.

[4] T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond Accuracy: Behavioral Testing of NLP Models with Checklist," arXiv preprint arXiv:2005.04118, 2020.

[5] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The Risk of Racial Bias in Hate Speech Detection," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668-1678.

[6] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional Tests for Hate Speech Detection Models," arXiv preprint arXiv:2012.15606, 2020.

[7] G. I. Sigurbergsson and L. Derczynski, "Offensive Language and Hate Speech Detection for Danish," in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 2020, pp. 3498-3508.

[8] A. K. Ojha, et al., "Detecting Insults in Social Commentary," available at: https://kaggle.com/c/detecting-insults-in-social-commentary.

[9] Y. Neuman, et al., "Metaphor Identification in Large Texts Corpora," PLoS ONE, vol. 8, no. 4, 2013, doi: 10.1371/journal.pone.0062343.