Detecting Hate Speech in Tweets with Advanced Machine Learning Techniques

J. David Sukeerthi Kumar[1], Dudekula Moulabi[2], Dudekula Hussainamma[2], Battula Kavya[2], Boothuru Veena Nissie[2] and Ganne Rakshitha[2]

[1]*Department of Computer Science & Design, Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India*

[2]*Department of Computer Science & Engineering, Santhiram Engineering College, Nandyal-518501, Andhra Pradesh, India*

Keywords: XGBoost, BERT, RoBERTa, Hate Speech Identification, Natural Language Processing.

Abstract: Twitter is a social media platform that enables the swift dissemination of information. Twitter users have the ability to broadcast brief messages called tweets. Tweets have news, opinions, and random thoughts. But they also have discriminatory messages, such as hate speech, that can promote hatred and hurt people as well as companies. Determining hate speech is key to maintaining a considerate and safe online community. The project aims to enhance internet security by identifying hate speech in tweets through machine learning techniques, specifically Natural Language Processing (NLP). Training of the model will be executed on a tagged dataset. The tweets can be preprocessed by eliminating special characters and stop words before being transformed into embeddings with the help of BERT/RoBERTa and XGBoost the proposed method significantly improves detection accuracy while reducing false positives, as measured by F1-score, recall, and precision.

1 INTRODUCTION

The fast development of social media sites such as Twitter has changed communication, but it has also enabled the propagation of harmful content, especially hate speech. Generally speaking, hate speech is any speech that spreads misleading information or disinformation and disparages, ridicules, or threatens people or groups because of their nationality, gender, ethnicity, religion, or other protected characteristics. Identifying hate speech is critical to making online environments safer, but the conventional keyword-based approach is not effective because hate speech is complicated and dynamic. Hate speech detection systems need to differentiate between hateful, offensive, and neutral speech and comprehend sarcasm, intent, and context. Conventional rule-based and machine learning-based methods like Logistic Regression, SVM, and Random Forest are less accurate when used with large datasets.

The study explores using deep machine learning techniques to identify hate speech on Twitter, focusing on deep learning architectures, natural language processing, and transformer architectures to understand contextual meaning. Figure 1 show the Hate Speech Detection Architecture.

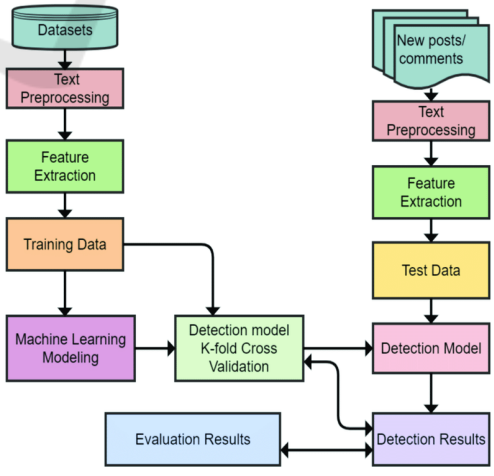

Figure 1: Hate speech detection architecture.

Kumar, J. D. S., Moulabi, D., Hussainamma, D., Kavya, B., Nissie, B. V. and Rakshitha, G.
Detecting Hate Speech in Tweets with Advanced Machine Learning Techniques.
DOI: 10.5220/0013923100004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 5*, pages 81-85
ISBN: 978-989-758-777-1

2 PROBLEM STATEMENT

This paper presents a high-precision hate speech detection system for Twitter, utilizing advanced machine learning techniques. The system uses data preprocessing, feature extraction, and contextual embeddings from transformer-based models. Performance is evaluated using precision, recall, and F1-score metrics, with hyperparameter search and ensemble learning enhancing performance.

3 LITERATURE REVIEW

Hate speech on social media, particularly Twitter, has been identified as a significant issue causing discrimination, cyberbullying, and disinformation in society. Traditional machine learning methods like Naïve Bayes (NB), Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF) were used in early hate speech detection models.

Davidson et al. (2017) proposed a hate speech corpus and utilized Logistic Regression and SVM for the purpose of classification and attained competitive outcomes. But classical models tend to face contextual understanding and generalization challenges, which renders them less efficient when it comes to identifying implicit hate speech.Next, Schmidt & Wiegand (2017)have surveyed hate speech detection with NLP methods by examining several feature engineering approaches (TF-IDF, n-grams, bag-of-words) and classification models (SVM, Naïve Bayes, Decision Trees). Fortuna & Nunes (2018) conducted a comprehensive study on hate speech detection in social media, highlighting dataset biases as a significant challenge and proposing fair classification mechanisms.

Researchers utilized deep learning techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to overcome the limitations of feature engineering.

While RNNs, specifically LSTM and GRU, capture long-distance dependencies in sequential text, CNNs find local structures in text. Badjatiya et al. (2017) proposed an LSTM-based approach to hate speech detection and reported improved results compared with traditional methods. Yoon (2016) examined CNN-based sentence classification and demonstrated that CNNs are powerful at detecting local patterns in text and thus useful for hate speech detection. LSTMs can handle sequential relationships and long-term dependencies more effectively than CNNs. Deep learning models, however, often lack interpretability and require large amounts of labeled data.

Recent advancements in Natural Language Processing (NLP) have led to the popularity of transformer-based models like BERT, RoBERTa, and ALBERT.After been pre-trained on sizable datasets and refined on hate speech corpora, these models use contextual embeddings to improve classification accuracy. Mozafari et al. (2020) utilized a transfer learning method using BERT to detect hate speech and demonstrated that, unlike CNNs and LSTMs, fine-tuning pre-trained transformers on hate speech corpora enhances precision and recall. The attention mechanism of transformers allows for end-to-end reasoning over verbal hints, such as implicit hate speech and sarcasm.

Hybrid approaches are methods that combine multiple architectures to improve classification accuracy. For instance, RoBERTa embeddings integrated with XGBoost (RoBERTa+XGBoost) and ALBERT integrated with BiLSTM (ALBERT+BiLSTM) both employ contextual embeddings and sequential modeling. Devlin et al. utilized the Bidirectional Encoder Representations from Transformers (BERT) model in bidirectional text processing (2018) to acquire contextual word representations. (Devlin et al. 2018). The method's understanding of sarcasm, implied hate, and intricate linguistic structures helped a great deal in increasing the precision of hate speech classification. Raffel et al. (2020) employed transformer models to explore transfer learning and highlighted the need for pre-training a model on large corpora and then fine-tuning it on a particular hate speech dataset. The research explained the challenges in scalability and generalization issues in hate speech classification.

The accuracy of hate speech identification has significantly increased as deep learning and transformer-based models have supplanted conventional machine learning techniques. Addressing bias and contextual understanding remains a primary research concern, notwithstanding the potential of hybrid and ensemble techniques. It will be necessary to combine ethical AI frameworks with developments in natural language processing to create more trustworthy hate speech detection systems.

4 EXISTING RESEARCH

The existing hate speech detection system discuss multiple techniques, including traditional classifiers, deep learning architectures.

4.1 Traditional Machine Learning Approaches

The existing approaches highlights the use of supervised learning methods that rely on feature engineering. The models used include:

- Support Vector Machines (SVM) – A widely used classifier for text-based hate speech detection.
- Naïve Bayes (NB) – A probabilistic model effective for handling textual data.
- Logistic Regression (LR) – Applied for binary classification tasks in hate speech detection.
- Random Forest (RF) – An ensemble learning technique that improves accuracy by aggregating multiple decision trees.
- Decision Trees (DT) – A rule-based method for classification.

These methods generally rely on TF-IDF, n-grams, and bag-of-words (BoW) for feature extraction. While they offer decent performance, they struggle with contextual understanding and fail to capture implicit hate speech.

4.2 Deep Learning-Based Approaches

To improve accuracy, the existing reasearchers discusses the application of deep learning models, including:

- Convolutional Neural Networks (CNNs) are utilized for feature extraction from text embeddings.
- (RNNs) are used for classification. – Captures sequential dependencies in tweets.
- Long Short-Term Memory (LSTM) – An advanced RNN variant that helps retain long-term dependencies.
- Gated Recurrent Units (GRUs) – A computationally efficient alternative to LSTMs with similar performance.

Models eliminate manual feature engineering and capture sequential patterns, but struggle with sarcasm and implicit hate speech.

4.3 Drawbacks in Existing System

- High False Positives & False Negatives – Most models incorrectly label non-hateful content as hate speech, resulting in high false positives.
- Bias in Training Data – Hate speech detection models tend to be trained on biased data, resulting in biased classifications across various demographic groups.
- Difficulty in Detecting Implicit Hate Speech – In contrast to explicit hate speech, implicit hate is context-dependent, sarcastic, and euphemistic, making it challenging for conventional models to identify.
- Computational Complexity – Transformer-based architectures such as BERT, RoBERTa, and ALBERT demand large computing power, thereby becoming hard to deploy for live hate speech identification.
- Overfitting in Deep Learning Models – Deep learning models like CNNs and RNNs tend to overfit on training data, and thus become futile for practical implementations.
- Scalability Problems – It requires highly scalable and efficient models to handle millions of tweets, posts, and comments in real-time, which is an issue in current systems.
- High Computational Cost – Deep learning models are computation intensive, which makes real-time hate speech detection challenging on low-power devices or on large social media sites.

5 PROPOSED SYSTEM

The proposed system consists of the following key components:

5.1 Data Collection & Pre-processing

- Collect tweets from publicly available datasets.
- Clean the data by removing URLs, hashtags, special characters, and stopwords.
- Apply tokenization and word embeddings.

5.1 Feature Extraction & Embedding

- Utilize transformer-based models (e.g., RoBERTa, ALBERT) for contextual word representations.
- Extract meaningful linguistic and semantic features to improve classification performance.

5.2 Model Implementation & Training

- Implement multiple state-of-the-art models and ensemble techniques for comparison.
- RoBERTa + XGBoost: Uses RoBERTa embeddings with XGBoost for classification.

- ALBERT + BiLSTM: ALBERT for feature extraction, followed by a BiLSTM model for sequence learning.
- Train models using labeled datasets with appropriate loss functions.

5.3 Evaluation & Optimization

- The comparison of models is based on metrics such as accuracy, precision, recall, and F1-score.
- Do hyperparameter tuning for improved performance.

5.4 Advantages of Proposed System

- Improved Accuracy: The algorithm uses sophisticated machine learning algorithms that enhance the precision of hate speech detection over standard approaches.
- Unification of Deep Learning: Implementation of deep learning models enables the system to learn contextual meaning in order to cut down on false positives and false negatives.
- Strong Feature Extraction: Higher order NLP principles assist in drawing strong features from linguistic and contextual features of tweets to enhance classification performance.
- Scalability: The system should be able to process large amounts of tweets without compromising performance, with consistent accuracy at high levels of data.
- Reduced Bias: The system leverages diverse datasets and algorithms to decrease bias in detection, resulting in fair and accurate classification.
- Far superior at identifying implicit hate speech and sarcasm than other machine learning models.

6 METHODOLOGY

- Data Collection: Collecting publicly available hate speech datasets (e.g.Davidson dataset).
- Data Cleaning & preprocessing – Converting all text to lowercase, removal of special characters, symbols, stop words and URLs, handle class imbalance using SMOTE (Synthetic Minority Over-Sampling Technique)**.**
- Feature Extraction: converting the textual data into numerical representations pretrained transformer models like RoBERTa and ALBERT.
- Model Selection & Training: Models like RoBERTa , XGBoost, and ALBERT are used to train the labelled dataset with optimization techniques such as hyper parameter tuning.
- Evaluation & Performance metrics: The trained models are assessed for their effectiveness in detecting hate speech using metrics like accuracy, precision, recall, and F1-score.

7 ARCHITECTURE

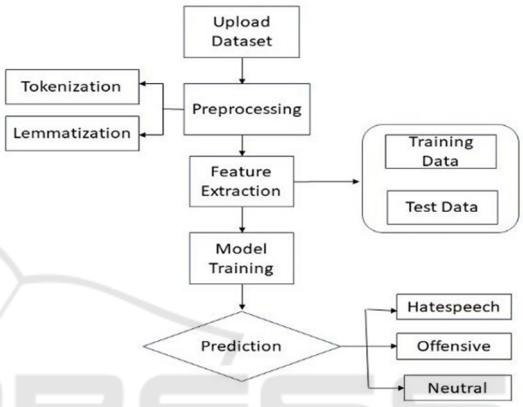

Figure 2: Design of the suggested system.

- **Uploading the dataset:** A publicly available Davidson hate speech dataset is gathered and uploaded in the software.
- **Data Preprocessing:** Preprocessing methods like tokenization, lemmatization is performed on tweets. Also, the data is preprocessed by eliminating stop words, and special characters.
- **Feature Extraction:** 80% of data is assigned for training from the uploaded dataset and 20% of data is assigned for testing. Raw data is then translated into numerical features.
- **Model Training:** To have high accuracy and interpretability, train the system RoBERTa and XGBoost models. Figure 2 show the Design of the suggested system.

8 RESULT

- Three categories, including hate speech, offensive speech, and neutral speech, will be used by the proposed method to classify tweets.
- High Accuracy and Improved Model Performance: Since we are using advanced

models like RoBERTa and XGBoost, the proposed system is expected to outperform traditional machine learning models.

- Enhanced Contextual Understanding: A notable decrease in false positives and false negatives, so that offensive but not hateful language is not incorrectly labelled as hate speech. Figure 3 show the Confusion matrix.
- Make use of social context: Rather than identifying single words in isolation, the model will examine conversational context to distinguish between hateful and non-hateful speech.
- Equitable datasets: Training datasets will be designed in a way to prevent biases against any particular demographic, political inclination, or ethnicity.

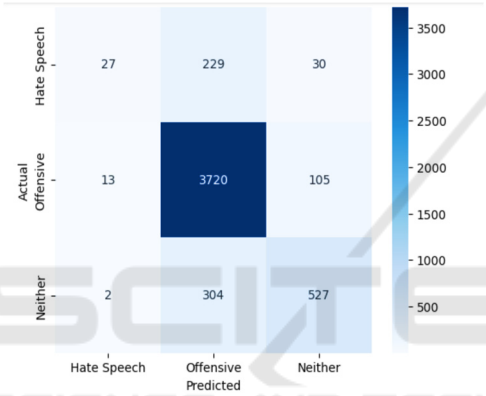

Figure 3: Confusion matrix.

9 CONCLUSIONS

The proposed system for hate speech detection on Twitter effectively combines advanced machine learning and transformer-based models like RoBERTa and XGBoost to achieve high accuracy and contextual sensitivity. Unlike traditional models, it significantly reduces false positives and false negatives by understanding nuanced language, sarcasm, and implicit hate. The system's design also emphasizes fairness by addressing dataset biases and ensuring demographic neutrality. With its scalable architecture and improved performance metrics, this approach is well-suited for real-time applications. Overall, it marks a step forward in building safer online platforms through intelligent content moderation.

REFERENCES

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). "Deep Learning for Hate SpeechDetection in Tweets". WWW.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). "Automated Hate Speech Detection and the Problem of Offensive Language". ICWSM.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). "Hate Speech and Offensive Language Dataset". Kaggle.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training ofDeep Bidirectional Transformers for Language Understanding". NAACL-HLT.

Fortuna, P., & Nunes, S. (2018). "Hate Speech Detection in Social Media: A Review".ACM Computing Surveys, 51(6), 1-30.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). "A BERT-Based Transfer LearningApproach for Hate Speech Detection in Online Social Media". ICMLA.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P. J. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". arXiv:1910.10683.

Schmidt, A., & Wiegand, M. (2017). "A Survey on Hate Speech Detection Using Natural Language Processing".Proceedings of the 5th International Workshop on Natural Language Processing for Social Media (NLP4SocialMedia).

Yoon, K. (2016). "Convolutional Neural Networks for Sentence Classification". EMNLP.