Text Analysis Using Pretrained Models for Hate Speech Detection

Rushi Parhad, Sharvayu Zade, Mukund Kuthe – Students, Symbiosis Institute of Technology, Nagpur
Dr. Deepak Suresh Asudani – Guide, Symbiosis Institute of Technology, Nagpur

INTRODUCTION

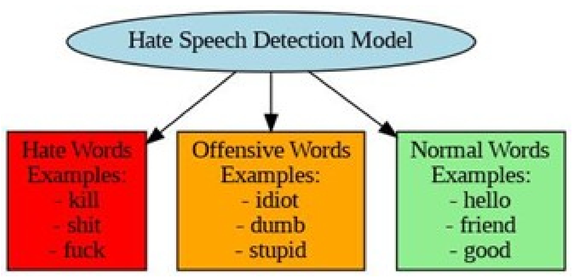
The rise in online hate speech presents severe challenges. This research develops a machine learning system for automated hate speech detection using models like XGBoost, SVM, and logistic regression trained on public datasets. Through TF-IDF and CountVectorizer, we transformed text into numeric form for training. Our results indicate XGBoost as the most accurate classifier, providing a solid foundation for moderation tools.

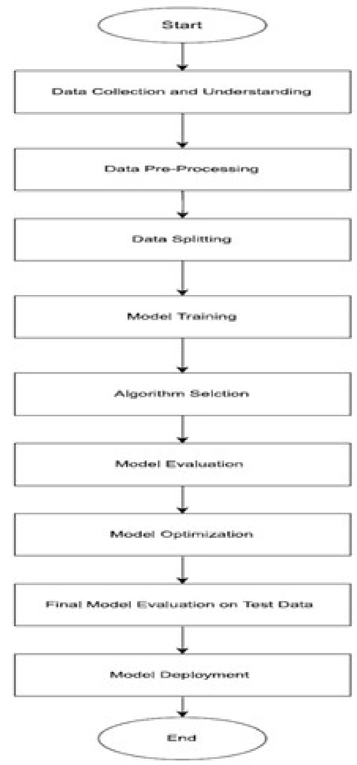
METHODS

1. Preprocessing Steps
Text is cleaned by removing URLs, mentions, and punctuation. All text is lowercased to ensure uniformity. Lemmatization (via WordNet) converts words to their base form. Stopwords are removed, but negations like “not” and “never” are kept to retain sentiment.

2. Vectorization Techniques
Text is converted into numbers using CountVectorizer and TF-IDF. Word2Vec and GloVe create dense word embeddings that capture context and word relationships, helping models understand meaning beyond word frequency.

3. Algorithms Used
Various machine learning models were applied to detect hate speech. Logistic Regression and Naive Bayes offer fast, simple baselines. Tree-based models like Decision Tree, Random Forest, and XGBoost handle complex patterns well. KNN provides instance-based learning. Among all, XGBoost delivered the highest accuracy.





DATASET AND ANALYSIS

A. Hate Speech and Offensive Language Dataset
This dataset contains 24,783 tweets, each labeled as either hate speech, offensive language, or neutral. It is widely used for text classification tasks and helps train models to detect harmful or inappropriate language. Labels are assigned by multiple annotators based on content related to race, religion, or sexual orientation.

B. ConvAbuseEMNLP Dataset
This dataset includes 12,768 annotated conversations between users and AI chatbots. Each message is labeled for different types of abuse like racism, sexism, or harassment. It captures both explicit and implicit hate speech in dialogue form, making it ideal for training conversational AI systems to recognize abusive behavior.

Analysis
1. XGBoost with TF-IDF achieved 0.9475 accuracy (80-20 split).
2. SVM performed best after XGBoost.
3. TF-IDF outperformed CountVectorizer in most scenarios.
4. SMOTE improved performance, especially for imbalanced datasets.

RESULTS

Model	TF-IDF 70-30	TF-IDF 80-20	CountVec 70-30	CountVec 80-20
Random Forest	0.943	0.9433	0.9411	0.9395
K Nearest Neighbours	0.9434	0.9439	0.942	0.9427
Support Vector Machine	0.9454	0.9463	0.9447	0.9439
XGBoost Classifier	0.9458	0.9482	0.9445	0.9473
Logistic Regression	0.945	0.9463	0.9439	0.9435
Decision Tree	0.926	0.9268	0.9262	0.9246
Naive Bayes	0.7541	0.7269	0.7543	0.7269

Table 1: Hate Speech Data With Smote

Model	CountVec 70-30	CountVec 80-20	TF-IDF 70-30	TF-IDF 80-20
Random Forest	0.9376	0.9379	0.9428	0.9421
K Nearest Neighbours	0.9388	0.9491	0.9422	0.9441
Support Vector Machine	0.9436	0.9441	0.9443	0.9445
XGBoost Classifier	0.9447	0.9465	0.9446	0.9475
Logistic Regression	0.94	0.9413	0.9442	0.9453
Decision Tree	0.9176	0.9163	0.926	0.9205
Naive Bayes	0.4757	0.4335	0.476	0.4343

Table 2: Hate Speech Data Without Smote

Hate Speech and Offensive Language Dataset

Model	CountVec 70-30	CountVec 80-20	TF-IDF 70-30	TF-IDF 80-20
Random Forest	0.7773	0.7807	0.8901	0.8935
K Nearest Neighbours	0.7862	0.7353	0.8389	0.8301
Support Vector Machine	0.7716	0.7823	0.8846	0.888
XGBoost Classifier	0.8721	0.8832	0.871	0.8253
Logistic Regression	0.8872	0.8865	0.8804	0.879
Decision Tree	0.7679	0.7772	0.882	0.8872
Naive Bayes	0.4667	0.4726	0.4811	0.4792

ConvAbuseEMNLPfull Dataset Without Smote

Model	CountVec 70-30	CountVec 80-20	TF-IDF 70-30	TF-IDF 80-20
Random Forest	0.8964	0.8943	0.8961	0.8947
K Nearest Neighbours	0.8917	0.8845	0.8804	0.8821
Support Vector Machine	0.905	0.9017	0.9042	0.9021
XGBoost Classifier	0.8912	0.89	0.8821	0.8891
Logistic Regression	0.893	0.8896	0.8883	0.8837
Decision Tree	0.8898	0.8876	0.8888	0.888
Naive Bayes	0.4644	0.4663	0.4704	0.471

ConvAbuseEMNLPfull Dataset With Smote

ConvAbuseEMNLP Dataset

CONCLUSION

This research demonstrates that XGBoost consistently outperforms traditional machine learning models in detecting hate speech, due to its capacity to capture complex patterns in high-dimensional text data. Careful preprocessing—including URL removal, lemmatization, and selective stopword filtering—and the use of TF-IDF vectorization were critical factors in achieving high accuracy. Additionally, the application of SMOTE to balance class distributions effectively mitigated bias toward majority classes, ensuring that minority hate-speech instances were properly learned.

```
Model loaded successfully!

Hate Speech Detection System
Type 'quit' to exit


==========================================
Input Text: hi how are you friend
==========================================

Prediction Results:
Predicted Class: Neither

Class Probabilities:
Hate Speech: 0.3038
Offensive Language: 0.2298
Neither: 0.4665

Key Words Influencing Prediction:
hi, friend
```

The output represents the execution of the final model, demonstrating its ability to analyze input text, calculate class probabilities, and highlight key words influencing the prediction.

ACKNOWLEDGEMENTS

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber
Paper: Automated hate speech detection and the problem of offensive language (2017)

Anna Curry, Glen Abercrombie, and Verena Rieser
Paper: ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI (2021)

Anna Schmidt and Michael Wiegand
Paper: A survey on hate speech detection using natural language processing (2017)

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen
Paper: BERTweet: A pre-trained language model for English tweets (2020)