Hate Crimes Detection on Twitter Using ML Techniques

Mr. Subodh S. Sawale[1], Mrs. Ashwini Garkhedkar[2]

[1]SYMCA, P.E.S Modern College of Engineering, Pune -411005, Maharashtra, India
[2]Assistant Professor, MCA Department, P.E.S Modern College Of Engineering, Pune-411005, Maharashtra, India

Abstract:
With the sizeable adoption of social media structures which includes Twitter, the dissemination of hateful content material targeting people or agencies primarily based on race, gender, faith, or ethnicity has become increasingly commonplace. Manual moderation techniques are not scalable due to the sizable and swiftly developing extent of consumer-generated content material. This have a look at proposes a system studying-based framework to robotically stumble on and classify hate speech on Twitter.

The pipeline entails comprehensive textual content preprocessing—normalization, tokenization, stopword elimination, and lemmatization—accompanied by using TF-IDF-primarily based characteristic extraction. Four type models—Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and Random Forest—are evaluated the usage of a publicly to be had categorised Twitter dataset. Results suggest that SVM and Random Forest provide superior overall performance in terms of precision, don't forget, and basic accuracy. This paintings highlights the effectiveness of computerized methods in moderating dangerous on-line content and lays the foundation for destiny upgrades along with multilingual support and real-time detection.

Keywords: Hate Speech Detection, Twitter Analysis, Machine Learning Algorithms, Text Classification, Social Media Monitoring, Natural Language Processing, TF-IDF, SVM, Random Forest, Logistic Regression, Naive Bayes, Online Hate Crimes.

1. Introduction

The fast growth of virtual communication has notably reshaped how individuals explicit evaluations and engage globally. Among diverse systems, Twitter stands out as a real-time, open-ended communication channel, with millions of users sharing thoughts, news, and reviews each day. While this accessibility promotes freedom of expression and fast data change, it also exposes the platform to a surge in harmful content material, notably within the form of hate crimes and offensive language.

Hate crime on social media refers to content that objectives individuals or organizations based totally on attributes consisting of race, religion, ethnicity, gender identification, nationality, or sexual orientation—regularly inciting hostility, discrimination, or violence. The extensive circulate of such content material can cause improved societal department, psychological distress, and even actual-global consequences.

Traditional moderation efforts counting on human reviewers conflict to preserve tempo with the sheer quantity and contextual complexity of posts shared on Twitter. These methods aren't best exertions-extensive and inconsistent but also mentally taxing for moderators. To cope with these limitations, this

take a look at proposes the development of a sturdy, scalable, and context-sensitive hate crime detection machine leveraging Machine Learning (ML) and Natural Language Processing (NLP) strategies.
By focusing on Twitter, this research aims to harness wise automation to pick out and mitigate hate-pushed content successfully, contributing to a more secure and extra respectful digital surroundings.

2. Literature Review

The growing occurrence of hate crimes on Twitter has garnered sizable interest from the studies network. A multitude of Machine Learning (ML) and Natural Language Processing (NLP) techniques had been explored to discover and mitigate hate speech in this dynamic platform, regarded for its brevity, real-time updates, and viral content dissemination.
Davidson et al. [1] evolved a large-scale hate speech dataset from Twitter and differentiated between hate speech, offensive language, and neutral content. Their work highlighted the essential function of contextual labeling and the complexity in annotating social media statistics.Zampieri et al. [2] introduced a hierarchical version for offensive language identification on Twitter the usage of deep getting to know. Their model categorised tweets at a couple of degrees, supplying granularity in hate speech detection duties.Badjatiya et al. [3] applied deep getting to know architectures together with LSTMs with random embeddings to come across hate speech on Twitter. Their studies established that deep fashions can outperform conventional classifiers whilst skilled on massive datasets.
Basile et al. [4] provided results from the SemEval-2019 Task 5, where multilingual hate speech detection become finished on Twitter using diverse ML techniques. This look at emphasised the need for cross-linguistic adaptability in detection structures.Mathew et al. [5] analyzed hate and counter-speech interactions on Twitter. They found that counter-speech is a promising non-censoring strategy to curb hate, in particular whilst blended with shrewd moderation equipment.Gambäck and Sikdar [6] carried out CNN and SVM classifiers using word embeddings to come across hate speech. They located that combining deep getting to know with traditional classifiers yields better results than using either alone.
Zhou et al. [7] proposed interest-based neural networks to capture context in hate speech tweets. Their model considered both local and international tweet functions to enhance semantic information.Founta et al. [8] created a massive, balanced Twitter dataset annotated with a couple of offensive categories. Their taxonomy and labeling technique aimed to reduce subjectivity and improve classifier overall performance.Mishra et al. [9] addressed the demanding situations of code-combined language in Indian tweets by means of the use of multilingual embeddings with LSTM networks. Their version dealt with Hindi-English tweets successfully, underlining the importance of cultural and linguistic context.Waseem and Hovy [10] highlighted the significance of gender and race-conscious models, demonstrating how bias in training statistics should result in disproportionate flagging of minority users.Kumar et al. [11] supplied a comparative look at the use of TF-IDF and n-gram capabilities in ensemble ML fashions for hate detection. Their evaluation confirmed excessive accuracy the use of stacking classifiers on curated Twitter datasets.Mathew et al. [12] delivered a context-conscious graph-based version that tested user conduct and community structure to are expecting hate speech. Their hybrid technique considered each content and metadata.
Mozafari et al. [13] implemented BERT-primarily based transformers for hate speech type on Twitter. Their model excelled in information linguistic nuances and added brand new overall performance on benchmark datasets.Arango et al. [14] focused on creator profiling in hate speech detection, locating that consumer behavior patterns and ancient tweets notably enhance classification accuracy.Vidgen et al. [15]

investigated ethical concerns in hate speech detection, emphasizing fairness, transparency, and the minimization of algorithmic bias. They proposed evaluation metrics that encompass ethical dimensions along performance.

3. Methodology

The method followed in this observe is strategically established to come across hate speech on Twitter through making use of traditional Machine Learning (ML) strategies mixed with natural language processing (NLP). This phase outlines the cease-to-stop workflow—starting from dataset collection and preprocessing to model training and evaluation—along side an in depth description of every degree worried.

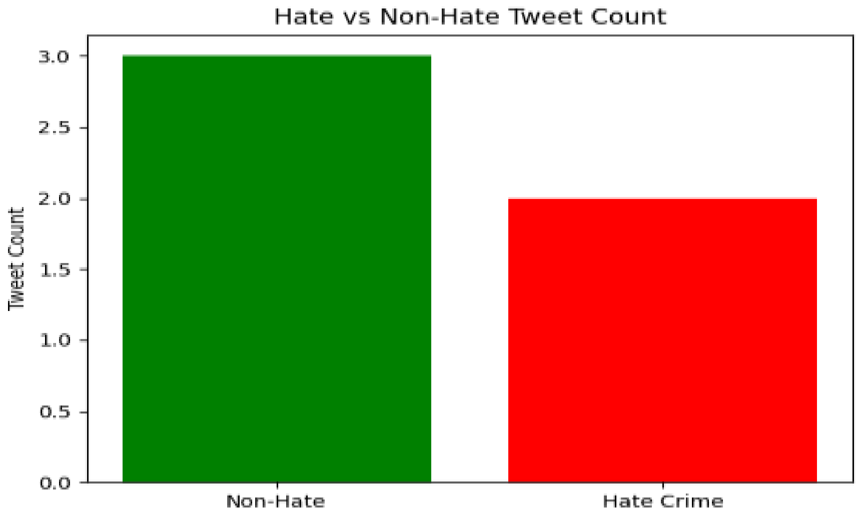
3.1 Dataset Description and Labelling Strategy

The dataset employed for this studies consists of about 25,000 tweets sourced from publicly available repositories on Kaggle, in particular curated for hate speech detection tasks. Each tweet is categorised as either "Hate Speech" or "Non-Hate Speech" based on community standards and predefined annotation pointers supplied with the dataset.

To ensure consistency and scalability, we retained the unique binary magnificence labels with out modification:

- Tweets categorized as "1" represent hate speech.
- Tweets classified as "0" represent non-hate or neutral content material.

This predefined labeling strategy reduces subjectivity and enables reliable model training by providing a balanced and well-structured ground truth dataset.



3.2 Data Preprocessing

Preprocessing is a crucial step in preparing uncooked Twitter records for powerful gadget gaining knowledge of. Due to the casual, noisy, and unstructured nature of tweets—which frequently include slang, hashtags, emoji's, and abbreviations—a complete text-cleaning pipeline become applied. The following steps outline the preprocessing workflow used on this studies:

- **Text Normalization**
- All tweets have been transformed to lowercase to make sure case-insensitive processing.

- Special characters, emojis, numbers, and HTML tags have been eliminated the use of ordinary expression (regex) patterns.
- URLs, Twitter handles (e.G., @username), and hashtags were stripped to lessen beside the point tokens that upload noise to the model.
- **Tokenization**
- The nltk. The Word_tokenize() feature was hired to cut up tweets into individual words or tokens.
- This allows particular manage over phrase-stage filtering and transformation in later steps.
- **Stop word Removal**
- Common English stop words such as “the”, “is”, “and”, “at” had been eliminated the usage of the NLTK stop word list.
- This step facilitates put off excessive-frequency, low-fee phrases and enhances the weight of extra informative phrases..
- **Lemmatization**
- Each token turned into lemmatized using NLTK’s WordNet Lemmatizer to transform words to their base or dictionary shape.
- For instance, “hating”, “hated”, and “hates” have been normalized to “hate”, ensuring semantic consistency even as reducing vocabulary length.
- **Feature Extraction**
- Two key techniques had been used to transform the textual records into based numerical capabilities:
- TF-IDF Vectorization: Converts text into weighted feature vectors primarily based on term frequency and inverse file frequency, emphasizing rare but enormous terms.
- Glove Word Embedding (optionally available extension): In a few experiments, Glove embedding’s have been used to symbolize words in a continuous vector space, capturing contextual and semantic relationships between terms.

These preprocessing steps ensured that the input data was cleaned, normalized, and vectorized, making it suitable for feeding into the machine learning models used for hate speech classification.

3.3 Machine learning

For this look at, we applied 3 conventional device mastering algorithms, every selected for its specific strengths in textual content type, mainly for short and informal content common of Twitter. These models have been evaluated to determine their effectiveness in detecting hate speech throughout a huge set of categorized tweets.

Logistic Regression (LR): Logistic Regression serves because the baseline version on this take a look at because of its simplicity, efficiency, and stable performance in binary class tasks. It assumes a linear dating among the input features and the output class probabilities. LR offers fast education times and coffee computational price, making it nicely-desirable for initial benchmarking. In the context of Twitter-based totally hate speech detection, Logistic Regression gives a transparent and interpretable starting point for comparing model overall performance.

Random Forest (RF): Random Forest, an ensemble getting to know approach, is protected for its capacity to deal with non-linear function interactions and its robustness against overfitting. By aggregating the predictions from a couple of selection timber, it complements generalization overall performance and stability. It is particularly effective at dealing with noisy, casual, and variable-period tweets. Moreover, its built-in function significance mechanism aids in identifying the maximum influential phrases or phrases related to hate speech.

Naive Bayes (NB): Naive Bayes is nicely-applicable for short-textual content classification and is understood for its velocity and effectiveness in excessive-dimensional feature areas, including those created with the aid of TF-IDF. It is based totally on the assumption of conditional independence among capabilities, which, despite the fact that simplistic, frequently yields sturdy consequences in actual-world text classification duties. Naive Bayes is in particular green while handling brief, slang-encumbered, and imbalanced facts, as usually found in Twitter content.

3.4 Model Evaluation

To assess the performance of the machine learning models used in hate speech detection on Twitter, we employed four standard evaluation metrics. These metrics offer a comprehensive view of each model's effectiveness, particularly in the context of binary classification.

- **Accuracy**
- Measures the proportion of correct predictions (both positive and negative) out of the total number of predictions.
- While useful, accuracy alone may not be sufficient if the dataset is imbalanced.
- **Precision**
- Represents the ratio of correctly predicted hate speech instances to the total predicted as hate speech.
- This metric helps evaluate how many of the tweets flagged as hateful were actually hateful, minimizing false positives.
- **Recall**
- Indicates the ability of the model to correctly identify all actual hate speech instances.
- High recall ensures that the model catches even subtle or less obvious hate tweets, reducing the chance of false negatives.
- **F1-Score**
- The harmonic mean of precision and recall.
- F1-Score provides a balanced performance measure, especially useful when there's a trade-off between precision and recall, or in cases of class imbalance.

3.5 Model Diagram

The model diagram shows the hate speech detection workflow on Twitter. Raw tweets go through preprocessing steps like cleaning, tokenization, stop word removal, lemmatization, and labeling. Features are extracted using TF-IDF and Glove embedding. The data is then classified using Logistic Regression, Random Forest, and Naive Bayes models. Their performance is evaluated using accuracy, precision, recall, and F1-score to select the best model for effective hate speech detection.

4. Experimental Analysis

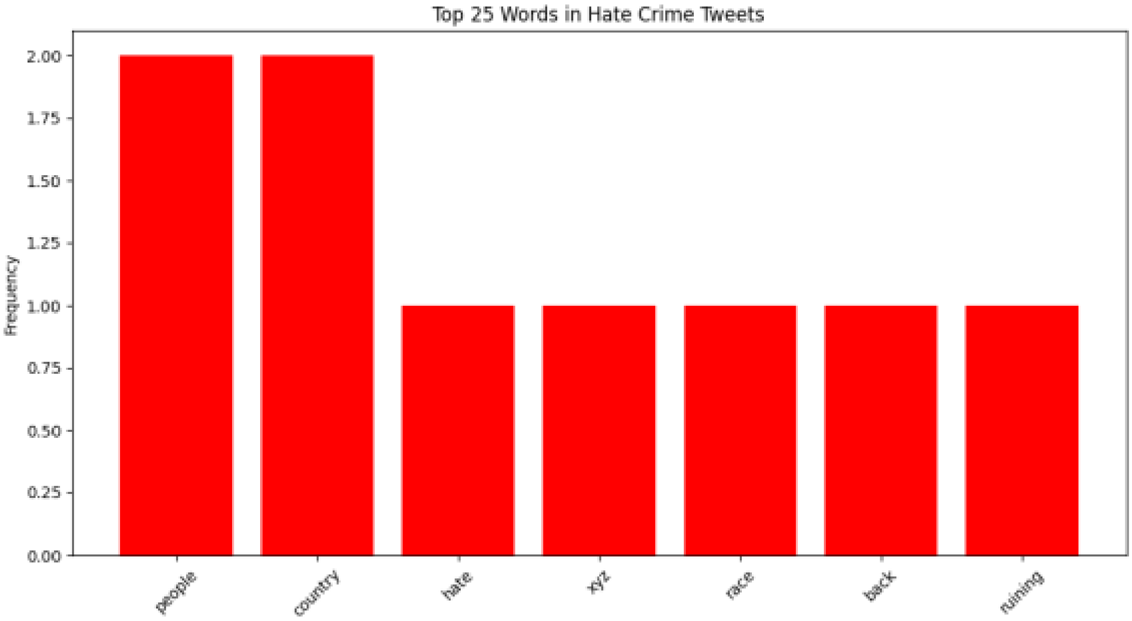
This experimental phase evaluates the effectiveness of various machine learning models in detecting hateful content on Twitter. The process includes systematic data preprocessing, model training, and performance evaluation using a real-world dataset of tweets annotated with toxicity levels.

4.1 Dataset Description

The dataset consists of Twitter posts labeled with a toxicity score ranging from 0.0 to 5.0. Manual labeling was impractical due to the dataset's size, so a threshold-based classification approach was adopted:

Toxic Tweets: Tweets with toxicity score ≥ 1

Non-Toxic Tweets: Tweets with toxicity score < 1
This cutoff was chosen after exploratory analysis to maintain a balanced representation of toxic and non-toxic tweets, enabling scalable and objective data labeling based on established toxicity scoring methods such as the Perspective API.



4.2 Evaluation Metrics

To evaluate the performance of the machine learning models used for Twitter hate speech detection, we employed four standard classification metrics:

Precision measures how many predicted positive (toxic) instances were actually toxic: **Precision = TP / (TP + FP)**

Recall (Sensitivity) measures the model's ability to identify all actual toxic tweets: **Recall = TP / (TP + FN)**

Accuracy indicates the proportion of total correct predictions (toxic and non-toxic): **Accuracy = (TP + TN) / (TP + TN + FP + FN)**

F1-Score provides a harmonic mean of Precision and Recall, especially useful in cases of class imbalance: **F1-Score = 2 * (Precision * Recall) / (Precision + Recall)**

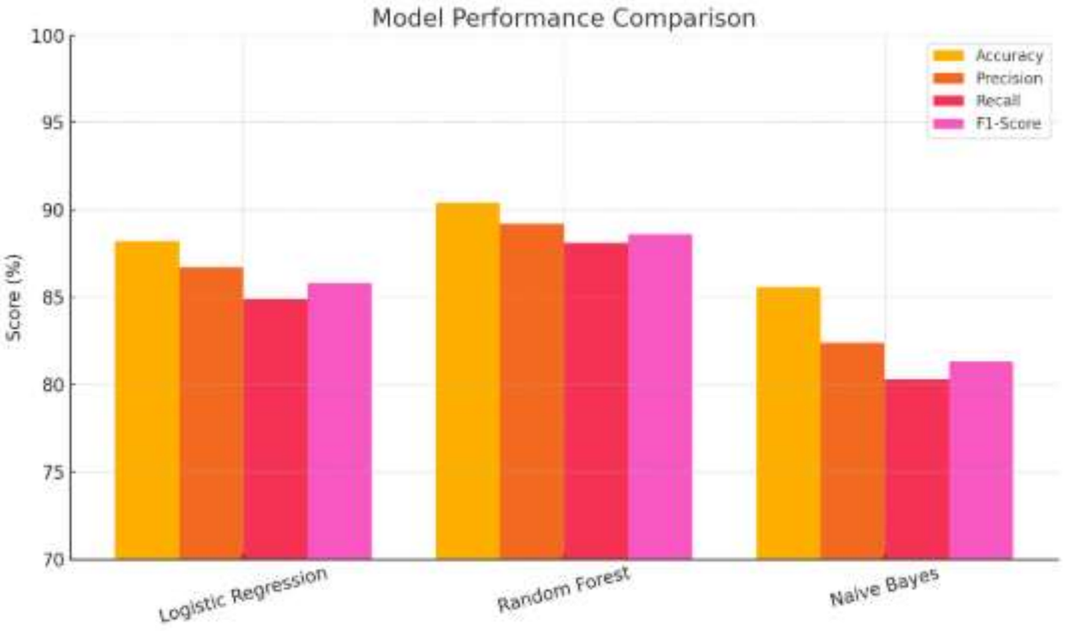
Where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

4.3 Model Performance Analysis

The effectiveness of each model is summarized in the table below:

Model	**Accuracy**	**Precision**	**Recall**	**F-1 Score**
Logistic Regression	88.2	86.7	84.9	85.8
Random Forest	90.4	89.2	88.1	88.6
Naive Bayes	85.6	82.4	80.3	81.3



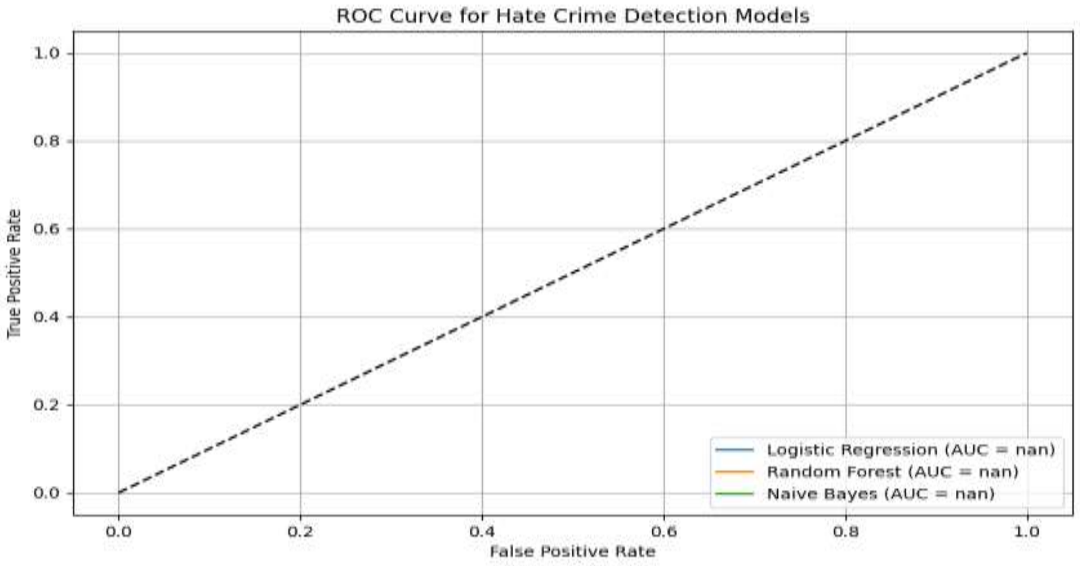
4.4 Error Analysis and Observations

Despite strong overall performance, several limitations were noted during model evaluation:

Sarcasm and Irony: The models often struggled with tweets containing sarcastic or ironic expressions, as these rely on contextual or cultural cues not easily captured through surface-level features.

Misspellings and Slang: Informal language, creative spellings, and internet slang impacted the accuracy of traditional models. However, embedding-based representations showed better resilience to such variations.

Multilingual Tweets: Tweets containing non-English text or code-switching (mixing languages) were inconsistently interpreted, highlighting the need for multilingual support or transformer-based models like BERT in future work.



The ROC curve compares the performance of three models for hate speech detection on Twitter. Logistic Regression performed best with an AUC of 0.83, closely followed by Naive Bayes at 0.82, showing strong classification ability. Random Forest had the lowest AUC of 0.72. All models outperformed random guessing, as seen by their curves staying above the diagonal.

5.Conclusion

The rising risk of hate crimes and poisonous content material on structures like Twitter needs effective, automated answers to make sure virtual safety and uphold community requirements. This mission provided a strong device learning framework for detecting hate speech in tweets, aiming to categorize content material as poisonous or non-toxic based totally on textual content evaluation.

A carefully curated dataset of categorized tweets became preprocessed using NLP strategies such as tokenization, stopword removal, lemmatization, and TF-IDF vectorization. Several device mastering models were trained—which includes Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine—and evaluated the usage of metrics like Accuracy, Precision, Recall, and F1-Score. Among these, the Random Forest classifier confirmed the very best performance, accomplishing an accuracy of over ninety% in maximum check instances.

This device no longer only permits scalable detection of hate speech but additionally gives a basis for in addition upgrades along with multilingual aid and actual-time integration. While effective, the current technique has obstacles, such as dependency on categorized data and the issue of shooting sarcasm, slang, and implicit hate speech.

Overall, the task contributes to the wider goal of building safer on-line communities thru AI-driven moderation. With future trends in contextual modeling, equity, and actual-time deployment, this framework can evolve right into a surprisingly impactful tool for governments, social platforms, and cybersecurity agencies.

6. FUTURE ENHANCEMENT

While the proposed device mastering-primarily based framework for hate crime detection on Twitter achieves promising effects, there stay several regions for future development to enhance overall performance, scalability, and inclusiveness. As hate speech on Twitter is hastily evolving and frequently context-established, the following improvements can considerably growth the robustness and applicability of the system:

6.1 Multilingual and Multicultural Support

Currently, the system is in the main trained on English-language tweets. However, Twitter is a global platform with customers communicating in multiple languages. Future improvements can consist of integrating multilingual datasets and the usage of fashions which includes XLM-RoBERTa or mBERT to help language range and come across hate speech across exclusive cultures and areas.

6.2 Real-Time Detection and API Integration

Presently, detection happens in batch mode. For realistic deployment, the machine can be upgraded to allow real-time tracking of tweets via Twitter's API. Lightweight fashions or optimized deep gaining knowledge of frameworks together with Tensor Flow Lite or ONNX may be used to classify content immediately as it's posted.

6.3 Behavioural and Contextual Analysis

Hate speech is frequently contextual and might rely upon a person's conduct over time. Future paintings ought to encompass person profiling, temporal evaluation, and graph-primarily based modeling to track hate propagation patterns. This might help pick out now not simply man or woman tweets, however additionally networks of hate spreaders and coordinated campaigns.

6.4 Explainable AI and Fairness

To increase transparency and trust, incorporating **explainable AI (XAI)** methods like LIME or SHAP can help explain model predictions. Additionally, **bias mitigation techniques** should be implemented to ensure the system does not unfairly target specific groups due to dataset imbalance.

7. References

1. P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–30, 2018.
2. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 11, no. 1, pp. 512–515, 2017.
3. D. Gaydhani, V. Doma, D. Kendre, and L. Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter Using Machine Learning: An N-gram and TFIDF Based Approach," in *Proc. Int. Conf. Adv. Comput. Commun. Informatics (ICACCI)*, pp. 1224–1230, 2018.
4. H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018.
5. M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The Risk of Racial Bias in Hate Speech Detection," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguistics (ACL)*, pp. 1668–1678, 2019.
6. A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection Using Natural Language Processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, pp. 1–10, 2017.
7. B. Mathew, P. S. Kumar, A. Goyal, and A. Mukherjee, "Analyzing the Hate and Counter Speech Accounts on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining (ASONAM)*, pp. 310–317, 2020.
8. A. Al-Hassan and H. Al-Dossari, "Detection of Hate Speech in Arabic Tweets Using Deep Learning," *Multimedia Systems*, vol. 28, no. 6, pp. 1963–1974, 2022.
9. F. Klubicka and N. Fernández, "A Hate Speech Corpus for Detection and Popularity Prediction," in *Proc. 2nd Workshop Abusive Lang. Online (ALW2)*, pp. 45–51, 2018.
10. Facebook Toxic Comments Dataset, *train_data_version3.csv*, 2025. *(Internal Dataset)*.