Detecting Racial and Ethnic Bias in Social Media Posts Using Machine Learning

Mira Camillo

Inspirit AI Scholars Program

19 June 2026

Abstract

Social media platforms contain large amounts of user-generated content, making it difficult for humans to identify harmful language manually. This project investigates whether machine learning can be used to detect racial and ethnic bias in social media posts. The dataset used was *Hate Speech and Bias against Asians, Blacks, Jews, Latines, and Muslims: A Dataset for Machine Learning and Text Analytics*, which contains 5,880 labeled tweets collected between 2020 and 2022. Several machine learning models were trained and evaluated using TF-IDF (Term Frequency-Inverse Document Frequency) text embeddings, including K-Nearest Neighbors (KNN), Random Forest, Multi-Layer Perceptron (MLP), AdaBoost, and Support Vector Classification (SVC). Random Forest achieved the highest accuracy (93.45%) and precision (57.14%), while the Multi-Layer Perceptron achieved the highest recall (11.54%) and F1 score (0.141). These findings suggest that machine learning can assist in detecting racial and ethnic bias in social media posts, although performance remains limited due to the imbalanced dataset and the complexity of language.

Introduction

Social media has become one of the most prominent ways people communicate, share ideas, and learn about the world around them. While these platforms can bring people together, they can also spread stereotypes, discrimination, and other forms of harmful content. As companies increasingly use machine learning to help identify biased content, it is important that these systems can recognize bias accurately and fairly. However, machine learning models are only as effective as the data they are trained on and the people involved in building them. If datasets do not include a wide range of experiences and perspectives, important forms of bias may be overlooked. Building better bias-detection systems therefore requires both diverse datasets and voices throughout the development process. This project investigates whether machine learning can effectively detect racial and ethnic bias in social media posts.

In this project, the term bias refers specifically to racial and ethnic bias directed towards Asians, Blacks, Jews, Latines, and Muslims. Tweets labeled as biased contain stereotypes, discriminatory language, or unfair generalizations about these groups. For example, a tweet that makes negative assumptions about a group of people based on their race or ethnicity would be considered biased. Because social media can influence how people think about others, it is important to identify this type of content accurately.

The huge amount of content posted on social media every day makes it difficult for people to review every message. Millions of posts are shared daily, making it nearly impossible for people to monitor everything that appears online. Machine learning offers a possible solution by automatically analyzing text and looking for patterns linked to biased language. If these systems can identify biased content accurately, they could help make online spaces safer and more welcoming for users.

This project is a classification problem because the model learns from examples that have already been labeled as biased or non-biased. The goal of this project is to determine how well different machine learning models can identify biased tweets and compare their performance. The models will be evaluated using accuracy, precision, recall, and F1 score to determine which approach is most effective for this task.

Background

Previous research has shown that machine learning can be used to identify harmful language on social media platforms. One study by Pereira-Kohatsu et al. (2019) developed a system called HaterNet to detect and monitor hate speech on Twitter. The researchers compared multiple learning models and found that preprocessing techniques such as TF-IDF improved performance by giving more importance to meaningful words while reducing the impact of common words. They also noted that social media language can be difficult to analyze because posts often contain slang, abbreviations and emojis.

Another study by Davidson et al. (2017) examined the challenge of distinguishing hate speech from offensive language on Twitter. The researchers found that models relying mainly on keywords often confused offensive language with hate speech. Their findings showed that context plays an important role when analyzing online language.

Together, these studies suggest that machine learning can be effective for identifying harmful content online, but performance depends on the quality of the data. These findings helped shape the methodology used in this project. Similar to these studies, this research uses TF-IDF features and multiple machine learning models to identify biased tweets and compare model

performance.

Dataset

The dataset used in this project was *Hate Speech and Bias against Asians, Blacks, Jews, Latines, and Muslims: A Dataset for Machine Learning and Text Analytics,* which was found through Google Dataset Search. The dataset contains 5,880 tweets and 7 columns. Each row corresponds to a single tweet, including replies, quotes and retweets.

The text analyzed in this project came from the Text column, which contains the full content of each tweet. Tweets were labeled as biased or non-biased by human annotators based on whether they contained racial or ethnic stereotypes or discriminatory language.

The dataset included the following columns: TweetID, Username, Text, CreateDate, Biased, Calling_Out, and Keyword. The main columns used in this project were Text, Biased, and Keyword. The dataset focuses on racial and ethnic bias directed toward Asians, Blacks, Jews, Latines, and Muslims.

Before training the models, the tweet text was preprocessed by converting words to lowercase, removing punctuation, removing common stopwords, and applying lemmatization. No tweets were removed during preprocessing, so the dataset contained 5,880 tweets both before and after preprocessing.

To convert the tweets into numerical features, TF-IDF (Term Frequency-Inverse Document Frequency) embeddings were used. TF-IDF gives importance to words that help distinguish between biased and non-biased tweets and reduces the influence of very common words.

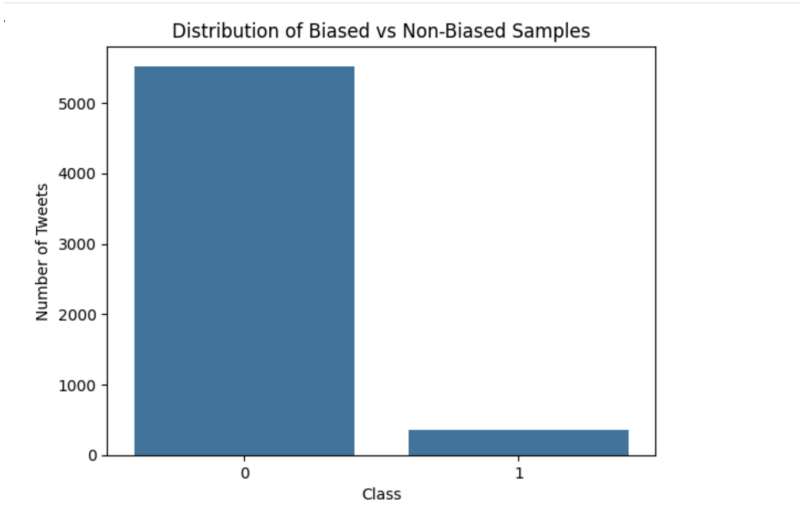

Figure 1. Distribution of Biased and Non-Biased Tweets.

Figure 1 shows that the dataset is highly imbalanced. Approximately 5,523 tweets are labeled as unbiased, while only 357 tweets are labeled as biased. Because biased tweets represent a small percentage of the dataset, accuracy alone is not sufficient for evaluating model performance.

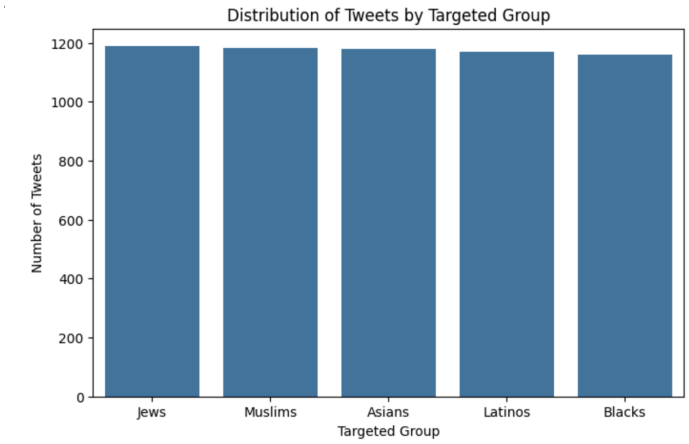

Figure 2. Distribution of Tweets by Targeted Demographic Group.

Figure 2 shows the distribution of tweets across the five targeted groups. The categories are relatively balanced, reducing the risk that one demographic group dominates the dataset.

Data Preprocessing

Social media text cannot be used directly by machine learning models because it often contains inconsistent punctuation, capitalization and other unnecessary text. Before training, several preprocessing steps were performed:

1. Converting all text to lowercase
2. Removing punctuation
3. Removing common stop words
4. Lemmatizing words

These steps helped make the text more consistent.

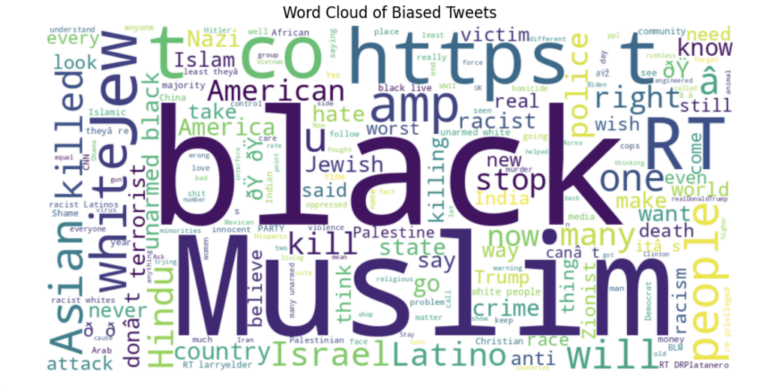

Figure 3. Word Cloud of Biased Tweets.

Figure 3 shows the most common words found in tweets labeled as biased. Many frequently appearing terms relate directly to the demographic groups represented in the dataset, which provides insight into patterns associated with biased content.

Methodology

The dataset was divided into training and testing sets using an 80/20 split. Approximately 80% of the tweets were used for training, while the remaining 20% were for testing.

The following machine learning models were tested:

Table 1. *Hyperparameters Used for Each Machine Learning Model*

Model	Hyperparameters
KNN	k = 3
Random Forest	N_estimators = 100, random_state = 42
MLP	Hidden_layer_sizes = (100,), max_iter = 300, random_state = 42
AdaBoost	N_estimators = 50
SVC	Default parameters

K-Nearest Neighbors (KNN) classifies tweets based on the labels of similar tweets that are closest to them. Random Forest combines multiple decision trees and uses voting

among them to make a final prediction. Multi-Layer Perceptron (MLP) is a neural network that can learn more complex patterns and relationships within text data. AdaBoost builds a series of models that focus on correcting mistakes made by earlier models to improve overall performance. Support Vector Classification (SVC) finds a boundary that best separates biased tweets from non-biased tweets.

The models were evaluated using four metrics:

- Accuracy: Percentage of all predictions that were correct.
- Precision: Percentage of tweets predicted as biased that were actually biased.
- Recall: Percentage of actual biased tweets that were correctly identified by the model.
- F1 score: A metric that combines precision and recall into a single value.

Recall is especially important in this project because the primary goal is identifying biased tweets. A model with low recall may miss many biased examples even if its overall accuracy appears high.

Results

The performance of each machine learning model is shown below:

Table 2. *Performance Metrics for Machine Learning Models*

Model	Accuracy	Precision	Recall	F1 Score
KNN	0.9269	0.2778	0.0641	0.1042
Random Forest	0.9345	0.5714	0.0513	0.0941
MLP	0.9065	0.1800	0.1154	0.1406
AdaBoost	0.9328	0.3333	0.0128	0.0247
SVC	0.9337	0.0000	0.0000	0.0000

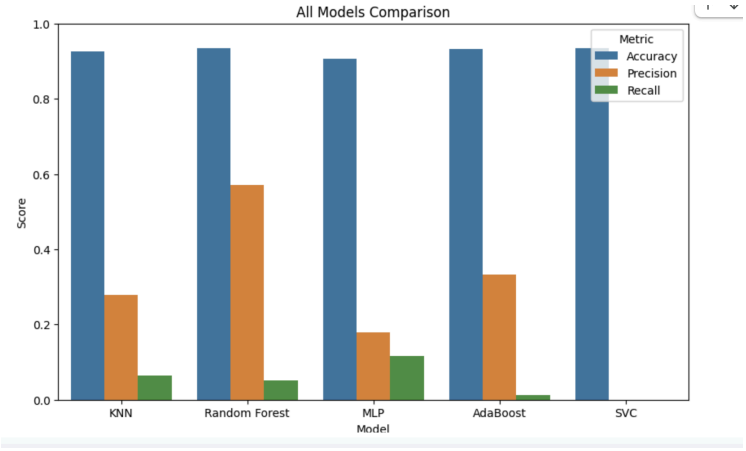

Figure 4. Comparison of Model Performance.

Figure 4 compares the accuracy, precision, and recall of all machine learning models.

Random Forest achieved the highest accuracy (93.45%) and the highest precision (57.14%). This indicates that when Random Forest predicted a tweet as biased, it was more likely to be correct than the other models.

However, because the primary goal of this project was to identify biased tweets, recall was particularly important. Recall measures the percentage of actual biased tweets that are correctly identified by the model. The Multi-Layer Perceptron achieved the highest recall (11.54%) and highest F1 score (0.141), making it the strongest model for bias detection despite having lower overall accuracy.

One surprising result was the performance of Support Vector Classification. Although SVC achieved an accuracy of 93.37%, it produced precision, recall, and F1 scores of zero. This shows that the model failed to identify any biased tweets and likely classified nearly all examples as non-biased. These findings further demonstrate why accuracy alone is not sufficient when evaluating models on imbalanced datasets.

Discussion

The results highlight the challenges of detecting racial and ethnic bias using machine learning.

One of the most important findings was that high accuracy did not necessarily mean strong bias detection. Although most models achieved accuracy scores above 90%, recall values remained low. This occurred because biased tweets made up only a small portion of the dataset.

Since most tweets were labeled as non-biased, models could achieve high accuracy by simply predicting the more common category.

This demonstrates why recall was especially important in this project. A model that misses biased content may appear accurate while still performing poorly at the main task.

Random Forest achieved the highest precision because it was more selective when identifying biased tweets. However, this approach reduced recall, causing it to miss many biased tweets.

The Multi-Layer Perceptron (MLP) achieved the highest recall and F1 score. This may be because neural networks can learn more complex patterns in text than some of the other models tested. Although its recall was relatively low, it identified more biased tweets than the other models.

The performance of SVC was also surprising. Despite achieving over 93% accuracy, it failed to identify any biased tweets. This demonstrates the importance of evaluating multiple metrics rather than relying on accuracy alone.

Future Work

Several improvements could help increase model performance in future studies.

First, methods for addressing the uneven distribution of tweets could be explored. Techniques such as increasing the number of biased tweets, reducing the number of non-biased tweets, or helping the model focus more on biased tweets may improve recall and identify more biased content.

Second, more advanced ways of representing text could be considered. While TF-IDF performed reasonably well, it does not fully capture the meaning of words within a sentence. Models such as BERT may be better at identifying subtle forms of bias because they consider the words around a term and the overall context of the tweet.

Finally, using larger and more diverse datasets could improve how well the models perform on new data and provide more examples of different types of biased language.

Conclusion

This project examined whether machine learning could be used to identify racial and ethnic bias in social media posts.

Using a dataset of 5,880 labeled tweets, several machine learning models were trained and evaluated using TF-IDF text representations. Random Forest achieved the highest accuracy and precision, while the Multi-Layer Perceptron achieved the highest recall and F1 score.

The results showed that machine learning can help identify biased tweets, although performance was affected by the uneven distribution of the data and the challenges of understanding language. While many biased tweets were correctly identified, others were still missed, highlighting the difficulty of detecting subtle forms of bias online.

Overall, the findings suggest that machine learning can serve as a useful tool for supporting bias detection on social media. Within this dataset, the best-performing models were able to identify biased content with encouraging results. Although there is still room for improvement, this study demonstrates the potential of machine learning to help address the spread of bias online. At the same time, creating fair and effective AI systems requires more than strong technical performance. As AI continues to play a larger role in shaping the content people see online, it is essential that individuals from diverse backgrounds have a seat at the table when these systems are built and evaluated.

Acknowledgements

I would like to express my sincere gratitude to my mentor, Matan Gans, for his guidance, feedback and support throughout this project. His suggestions helped strengthen both the methodology and the overall quality of this research paper. I would also like to thank Inspirit AI for providing the educational resources and learning environment that made this research possible.

References

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language* *. https://arxiv.org/pdf/1703.04009

Gunther Jikeli, Sameer Karali, & Soemer, K. (2023). Hate Speech and Bias against Asians, Blacks, Jews, Latines, and Muslims: A Dataset for Machine Learning and Text Analytics. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.8147308

NLTK. (2009). *Natural Language Toolkit — NLTK 3.4.4 documentation*. Nltk.org. https://www.nltk.org/

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. *Sensors*, *19*(21), 4654. https://doi.org/10.3390/s19214654

Scikit-learn. (2024). *scikit-learn: Machine Learning in Python*. Scikit-Learn.org. https://scikit-learn.org/stable/