

# Inferring Hate Speech Trends for Contemporary Tweets Using a Novel Machine Learning Approach from Supervised Learning Algorithms

Aryan Singhal

## Abstract

Social media platforms such as Twitter have become ubiquitous in our contemporary society, providing a platform for individuals to express their opinions and engage in discussions on a wide range of topics including those that are neutral and controversial. However, the growing popularity of Twitter has also led to an increase in the prevalence of hate speech, which raises concerns about its impact on individuals and society. This research investigates hate speech trends on Twitter by utilizing supervised machine learning classification, specifically Naive Bayes, and employing Natural Language Processing (NLP) features such as on a range of neutral and controversial topics. The study compares the prevalence of hate speech in these topics and tracks such trends from January 2022 to January 2023. The results show that hate speech was nearly 400% more prevalent in controversial topics than in neutral topics over the course of the year. In addition, this research finds that controversial topics are consistently more vulnerable to hate speech throughout the course of the year when compared to neutral topics. To conduct this study, a Multinomial Naive Bayes classification model was trained on a publicly available Twitter dataset that was specifically labeled for semantic hate speech and achieved an accuracy rate of 94.46%. Ultimately, the higher vulnerability of controversial topics should necessitate policymakers to introduce stricter warnings or frequent policy reminders to platform users. Such changes will foster a respectful and inclusive platform for users, preserving their freedom of expression and encouraging constructive discussions.

## 1. Introduction

Social media has become an integral part of our lives, providing a platform for people to connect, share ideas, and form communities. However, the increasing prevalence of hate speech and microaggressions on these platforms has become a cause for concern. Hate speech is any language or conduct intended to demean, degrade, or dehumanize a group of people based on

their race, ethnicity, gender, sexual orientation, religion, or other characteristics. Such speech not only harms individuals but also creates a toxic environment that discourages healthy discussions.

To combat this problem, researchers have turned to machine learning and natural language processing (NLP) techniques to analyze the prevalence of negative sentiments in tweets [10][18]. By quantifying the prevalence of hate speech, policies can be put in place to create a respectful and inclusive environment for everyone.

## **2. Research Context**

In the era of social media dominance, the analysis of sentiments expressed in tweets has garnered considerable attention from researchers. The impact of hate speech on online platforms, particularly on Twitter, has become a subject of concern due to its potential to propagate discrimination and incite violence. Numerous studies [12] have delved into the development and evaluation of machine learning (ML) models for sentiment analysis of tweets, aiming to accurately classify the sentiment conveyed within these short textual messages.

To enhance the performance of sentiment analysis models, researchers have employed various natural language processing (NLP) techniques for feature extraction. These techniques include tokenization, lemmatization, stemming, and the creation of unigrams, as well as part-of-speech tagging [10]. Additionally, researchers have curated and made available publicly labeled Twitter datasets specifically tailored for sentiment analysis tasks, facilitating the development and evaluation of ML models [12].

While several ML models, such as LinearSVC, AdaBoostClassifier, and LogisticRegression, have demonstrated comparable performance in sentiment analysis tasks, the Multinomial Naive Bayes classifier has emerged as a particularly effective choice for detecting hate speech on Twitter. Its ability to handle text data with multiple features and its capacity to capture the underlying probability distribution of words make it a well-suited approach for identifying instances of hate speech within tweets.

Despite the extensive research on sentiment analysis in tweets, a critical gap remains in understanding the prevalence of hate speech across controversial and neutral topics on Twitter, and how this prevalence has evolved over the past year. Previous studies have primarily focused on general sentiment analysis or targeted specific domains, overlooking the investigation of hate speech trends concerning the nature of the topics being discussed.

This study aims to fill this research gap by examining the prevalence of hate speech in contemporary tweets, distinguishing between controversial and neutral topics. By analyzing a

large corpus of tweets collected over the past year, this research seeks to shed light on the changing landscape of hate speech on Twitter and its association with different types of discussions. Through the application of a novel machine learning approach specifically designed to infer hate speech trends, this study will contribute to a deeper understanding of the dynamics of hate speech in the context of social media discourse.

### 3. Objective

The objective of this study is to develop a hate-speech detection model utilizing the Naive-Bayes Classifier, a highly effective supervised machine-learning algorithm. The model will be trained to accurately classify tweets based on their occurrence of hate speech, as well as group them by topic and keyword over monthly periods within the last year. Through the classification process, trends in hate speech can be identified and analyzed to determine the impact of political, social, and economic events on the well-being of society.

The examination of hate speech trends is crucial in gaining valuable insights into the evolution of social norms and ethical standards in recent times. Through the identification of these trends, effective measures such as new internet policies, laws, and enforcement strategies can be developed to guarantee a safer and more inclusive online milieu. This study endeavors to make a substantial contribution towards this critical objective by presenting an all-encompassing analysis of hate speech trends and proposing strategies for its mitigation, thereby preserving a respectful platform.

### 4. Materials and Methods

This study is performed with the following procedure.

**Step 1:** Identify a representative sample of controversial and neutral topics.

**Step 2:** For each topic, scrape 10,000 tweets using the SNScrape library [8], for every month in the year spanning 2022 to 2023.

**Step 3:** Train a Multinomial Naive Bayes model from a labeled Kaggle dataset [26] for hate speech (supervised learning)

**Step 4:** Use the trained model to predict the sentiment of the scraped tweets and examine the accuracy and recall for the predicting model

## 1. Representative Topics

To identify the top contemporary topics discussed extensively on social platforms, a comprehensive web search using Google search results was conducted. The aim was to select topics that are both neutral and controversial, ensuring a well-rounded representation. Table 1 showcases the chosen topics, with five being selected from neutral discussions and another five from controversial conversations. The bolded topics in Table 1 were utilized as search keywords during the process of scraping tweets for analysis.

Neutral	Controversial
<b>“food”</b>	<b>“meat”</b>
<b>“health”</b>	<b>“crisis”</b>
<b>“humanities”</b>	<b>“rights”</b>
<b>“weather”</b>	<b>“global warming”</b>
<b>“politic”</b>	<b>“trump”</b>
<b>“sports”</b>	<b>“conflict”</b>
<b>“hobbies”</b>	<b>“reform”</b>
<b>“finance”</b>	<b>“wage”</b>

Table 1: A sampled set of top neutral and controversial topics, garnered from the web.

## 2. Tweet Collection and Scraping

To gather a comprehensive dataset for analysis, tweets were scraped utilizing the publicly available Twitter Snsrape library [8]. The scraping process targeted each of the bolded topics listed in Table 1, treating them as keywords. Specifically, 10,000 English tweets were scraped for each month, spanning from January 2022 to January 2023.

The data flow of the scraped tweets is illustrated in Figure 1. Initially, the tweets were retrieved from the Twitter Database and saved as a .csv (comma-separated values) data table file in a local Google Colaboratory notebook. Subsequently, the data was transferred to Google Drive for long-term storage, ensuring accessibility for future prediction and analysis.

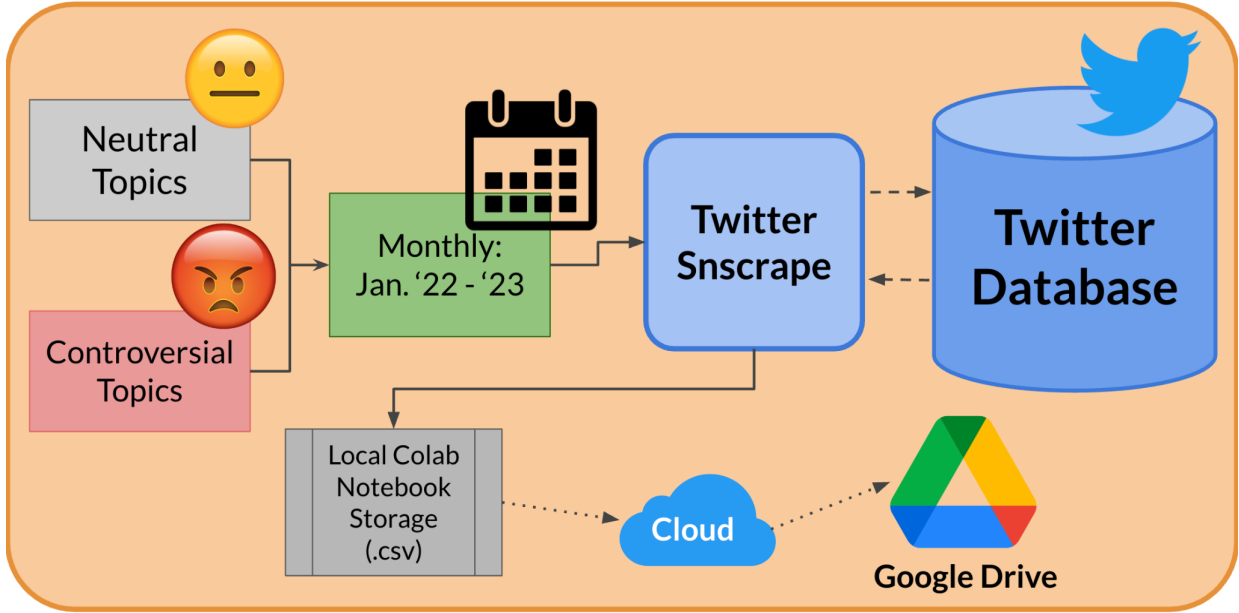


Figure 1: Overview of the entire dataflow, highlighting the journey of the monthly scraped tweets from the Twitter Database to the local Google Colaboratory notebook and finally to Google Drive for storage purposes.

### 3. Building and Training the Model

In this study, a supervised learning approach is employed to predict hate speech trends in contemporary tweets. To build the training model, a labeled dataset consisting of over 32,000 English tweets was obtained from Kaggle [26]. These tweets were prelabeled by expert linguists in a binary manner, where a value of '1' indicates the presence of racist or sexist content, while '0' denotes the absence of such content.

Prior research has demonstrated the effectiveness of machine learning techniques, such as Support Vector Machines (SVM) and Naive Bayes, for opinion mining [12]. In this study, a Multinomial Naive Bayes model is constructed. Initially, the dataset is processed to remove null values and eliminate duplicate entries. Subsequently, the labeled dataset undergoes cleaning procedures utilizing Natural Language Processing (NLP) techniques, including punctuation and stop word removal, as well as tokenization. This prepares a refined model training set for further analysis.

To facilitate machine learning, the cleaned text data is transformed into a matrix of token counts for each row value in the dataset, employing Python's CountVectorizer (CV) class. Naive Bayes classification is then applied to the transformed matrix. The dataset is split into a training set and

a test set using an 80%-20% ratio, with the test set comprising 20% of the data. This split is performed on both the features and labels of the matrix. The training data, derived from the split matrix, is utilized to fit and construct a Naive Bayes model, incorporating the corresponding labels.

Subsequently, the constructed model is evaluated on the reserved test set. A confusion matrix (Figure 3) is computed and generated to assess the accuracy of the model. This matrix provides a comprehensive representation of the model's predictive performance.

Figure 2 below demonstrates the flow of data when training the model.

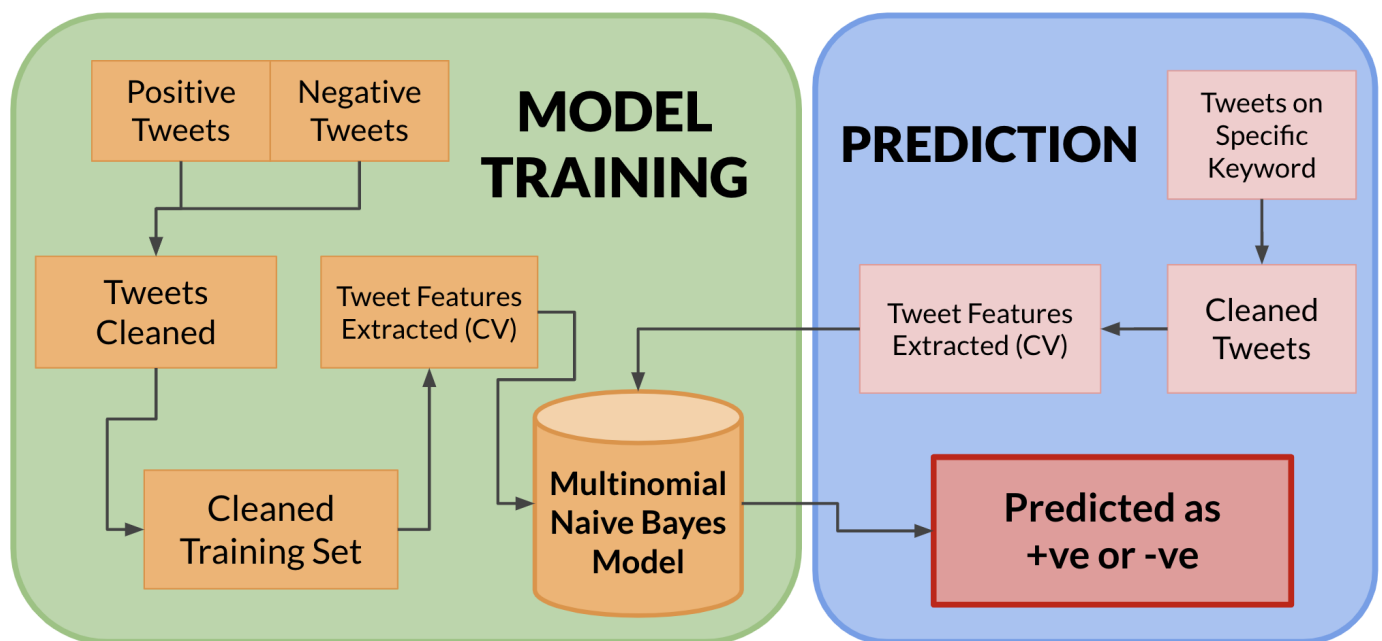


Figure 2: Overview of data flow for training model and sentiment prediction methodology.

#### 4. Predictions and Analysis

The scraped tweets collected in Step 2 undergo a similar process, as depicted in Figure 2. Each .csv file containing the tweets is subjected to cleaning and feature extraction. These cleaned and extracted features are then utilized with the previously trained and built Multinomial Naive Bayes model to predict sentiment.

The classification report provides a comprehensive evaluation of the model's performance. It includes metrics such as precision, recall, F1 score, and support. Precision represents the proportion of correctly predicted positive instances out of all instances predicted as positive.

Recall, also known as sensitivity, measures the proportion of correctly predicted positive instances out of all actual positive instances. The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's accuracy. Support indicates the number of instances in each class.

An in-depth look into the classification report:

- **Accuracy:** The overall accuracy of the model is 0.94, indicating that it correctly predicts sentiment in 94% of cases.
- **Macro Average:** The macro average of precision, recall, and F1 score is 0.77, indicating a decent overall performance across the classes.
- **Weighted Average:** The weighted average of precision, recall, and F1 score is also 0.77, suggesting that the model performs consistently across the classes, considering the class distribution.
- **Precision, Recall, and F1 Score:** The precision and recall for class 0 (negative sentiment) are 0.97 and 0.97, respectively. For class 1 (positive sentiment), the precision and recall are 0.57 each. The F1 score represents the harmonic mean of precision and recall and is also provided for each class.
- **Support:** The support indicates the number of instances in each class. There are 5,972 instances in class 0 and 421 instances in class 1, out of a total of 6,393 instances.

Figure 3a presents the precision and recall values obtained from the predictions made by the Multinomial Naive Bayes model. Figure 3b illustrates the confusion matrix, which showcases the accuracy of the model in classifying instances into their respective sentiment categories.

	precision	recall	f1-score	support
0	0.97	0.97	0.97	5972
1	0.57	0.57	0.57	421
accuracy			0.94	6393
macro avg	0.77	0.77	0.77	6393
weighted avg	0.94	0.94	0.94	6393

Figure 3a: Precision and recall of the predictions from the Multinomial Naive Bayes model.

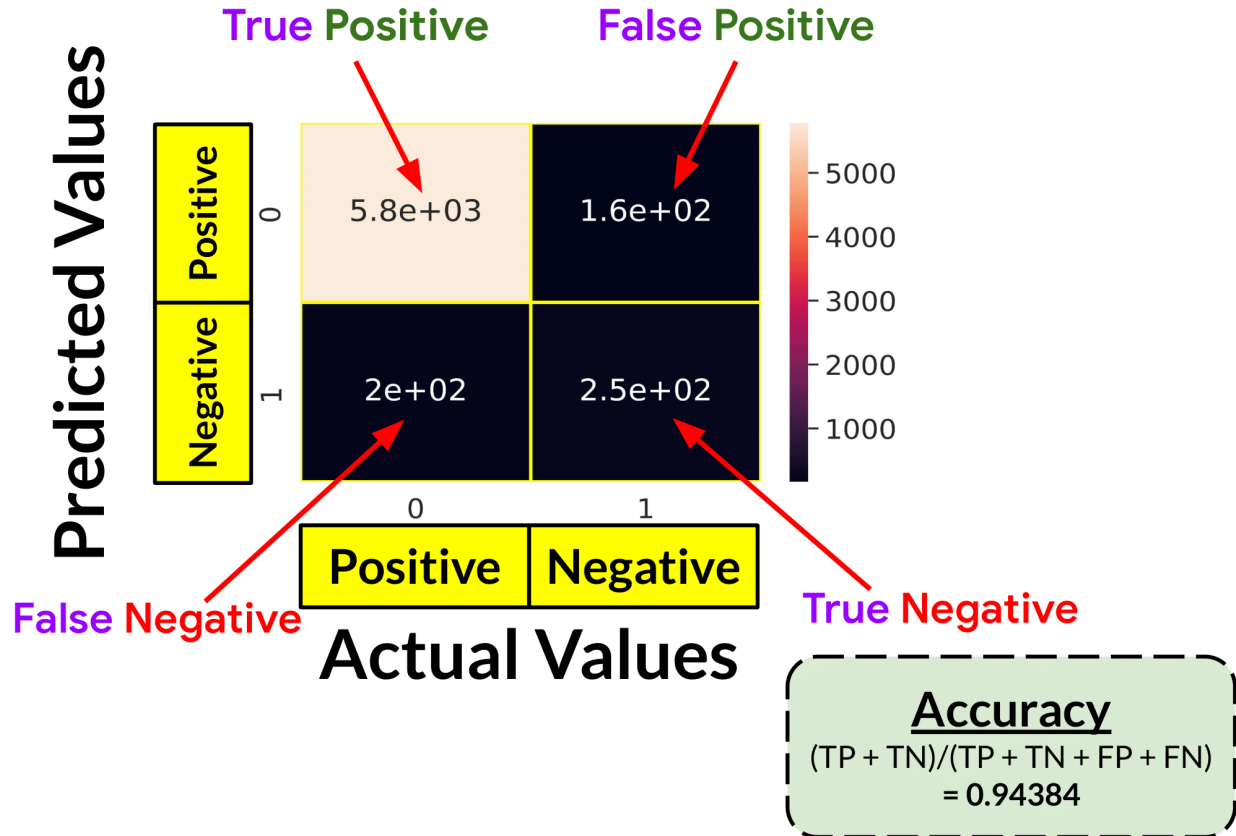


Figure 3b: The confusion matrix displaying the accuracy of the model.

## 5. Results and Observations

The distribution of negative tweets, as depicted in Figure 4, highlights an intriguing finding. From January 2022 to January 2023, it is evident that negative tweets are significantly more prevalent when discussing controversial topics compared to neutral ones. The analysis reveals that negative tweets on controversial topics are approximately four hundred times more likely to occur than on neutral topics.



Number of Hate Related Tweets for Neutral and Controversial Topics

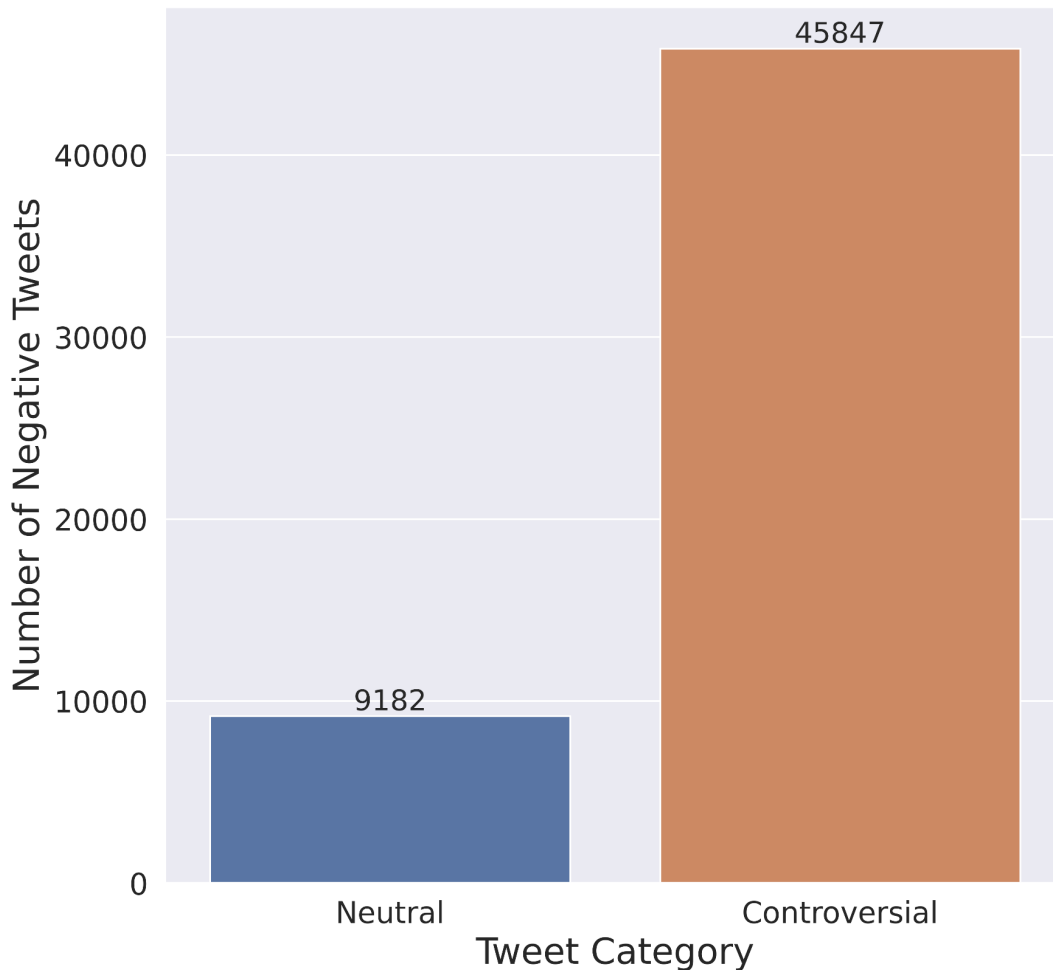


Figure 4: Distribution of negative tweets aggregated across neutral V.S. controversial topics

Figure 5 and Figure 6 delve deeper into the percentage of tweets predicted as negative for both neutral and controversial topics. Notably, these graphs demonstrate that the percentage of negative tweets remains relatively consistent month over month for each topic category throughout the observed period. This consistency suggests that there has been no drastic improvement or deterioration in the prevalence of hate speech within these topics from January 2022 to January 2023.

Moreover, it is worth noting that while the percentage of negative tweets for neutral topics mostly remains below the overall negative percentage line, controversial topics consistently approach or surpass the overall line. This stark contrast indicates that controversial topics are particularly susceptible to hate speech and underscores the necessity for heightened scrutiny, revisions in social platform policies, and effective enforcement strategies.

Figure 5 showcases the percentage of negative tweets specifically related to neutral topics, while Figure 6 displays the percentage of negative tweets for controversial topics. The dash-dot line in both figures represents the overall percentage of negative tweets across both categories of topics.

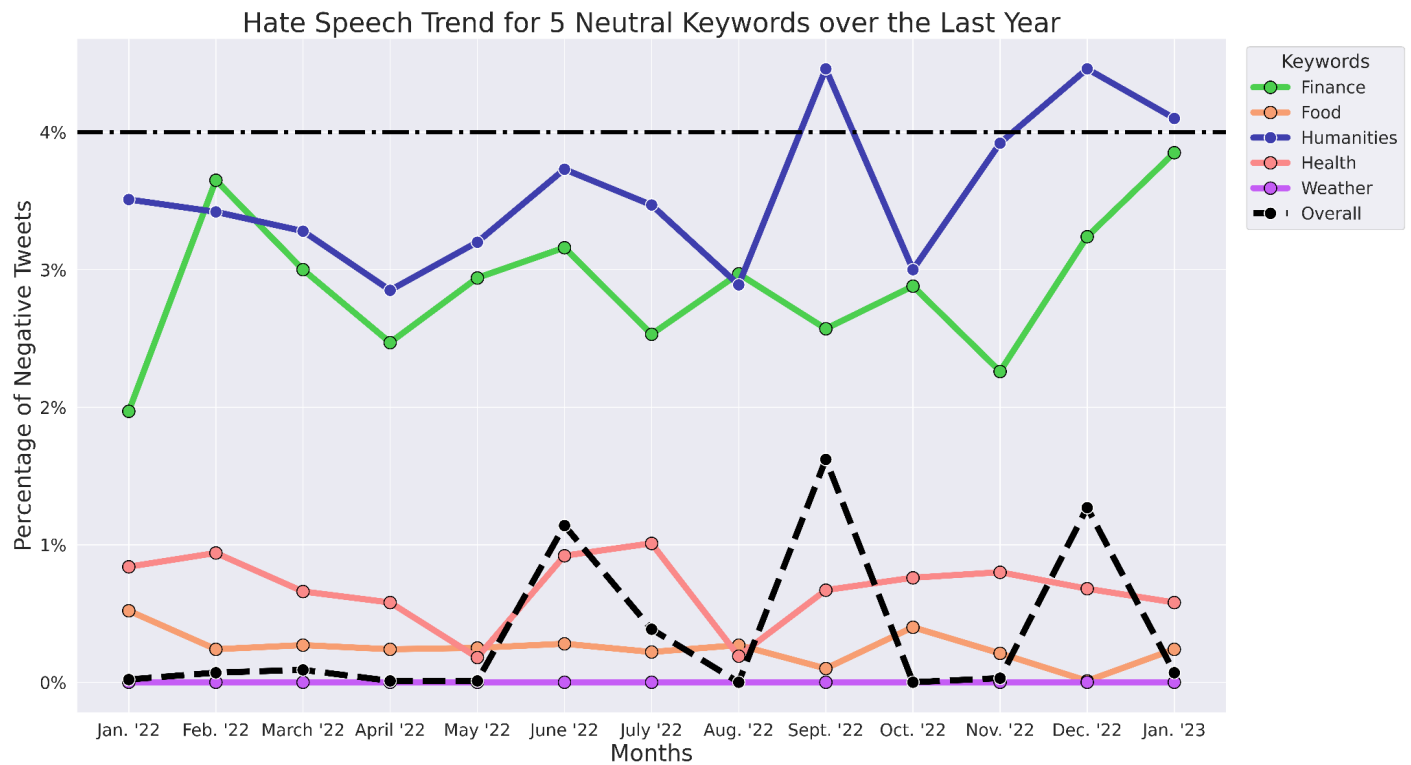


Figure 5: Hate speech trend quantified by percentage for negatively labeled tweets on selected neutral topics (keywords) from January 2022 to January 2023.

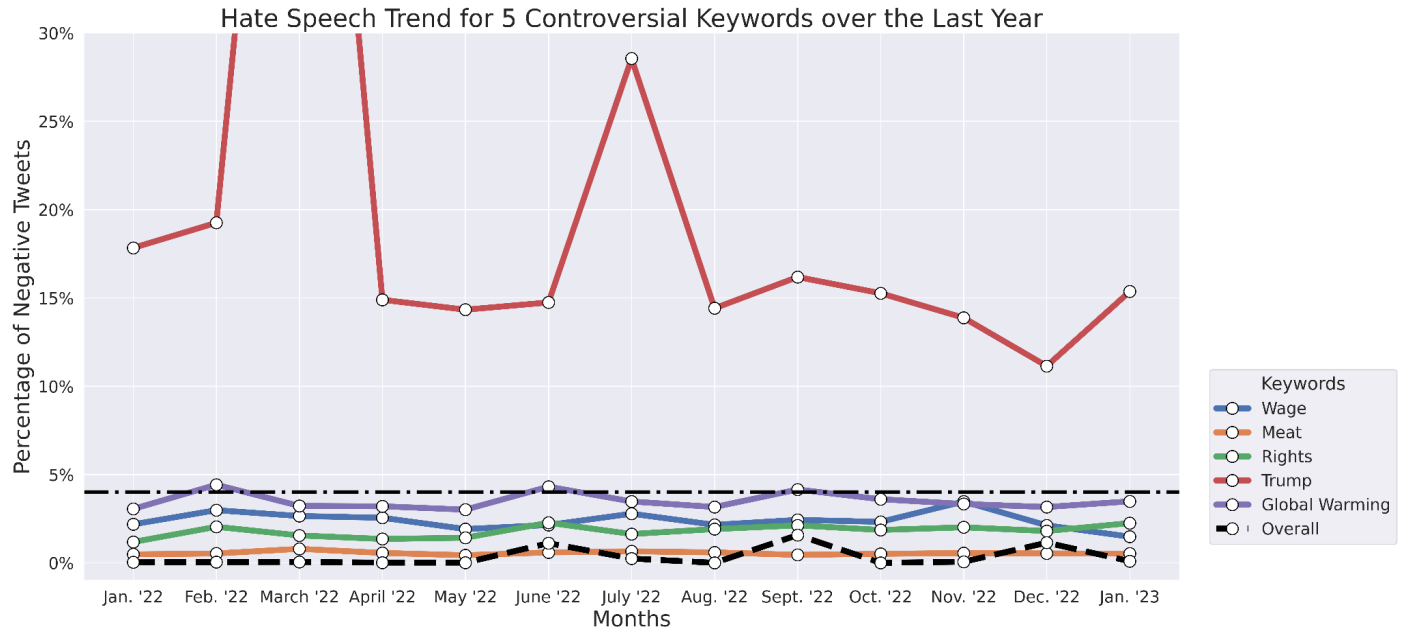


Figure 6: Hate speech trend quantified by percentage for negatively labeled tweets on selected controversial topics (keywords) from January 2022 to January 2023.

## 6. Conclusions

In this study, machine learning predictions revealed that hate speech is significantly more prevalent in discussions on controversial topics, four hundred times more likely compared to neutral topics. Moreover, the prevalence of hate speech remained consistent throughout the 2022-2023 period.

To address this persistent issue, social platforms must proactively revise and enforce their policies, striking a balance between fostering healthier discussions and preserving freedom of expression. Stricter policies targeting controversial topics and regular interstitial policy reminders can be effective strategies.

By taking these measures and promoting responsible engagement, social platforms can contribute to creating a safer and more inclusive online space. Future research should focus on exploring innovative machine-learning techniques to further mitigate hate speech and encourage constructive digital conversations.


## 7. Future Work

While this study has provided valuable insights into the trends of hate speech on Twitter, several avenues for future research would enhance the current findings. Below is a list of areas for further exploration.

1. Comparisons between hate speech trends in the USA and the rest of the world: This study focused on hate speech trends in the USA, but comparing trends between the USA and the rest of the world would reveal more. This could help policymakers and social media companies understand the prevalence and nature of hate speech across different regions.
2. Expanding features beyond n-grams: This study used n-grams to classify tweets, but semantic similarity measures could improve accuracy. Incorporating these features could enhance accuracy and understanding.
3. Experimenting with a broader range of topics: This study analyzed hate speech trends for 10 topics. Other topics, such as gender, sexual orientation, and disability, could be explored. This would provide insights into hate speech across marginalized communities.
4. Classifying negative tweets into niche categories: This study classified tweets as neutral or controversial, but negative tweets can be further categorized. Using machine learning to classify negative tweets could help policymakers and social media platforms to better understand and police negative content.

The future work outlined above has the potential to build on this study and provide a more comprehensive understanding of hate speech trends on Twitter. The results of this study may inspire further research and the development of more effective strategies for combating hate speech online.

## 8. References

- [1] Alessandro RussoAlessandro Russo 17311 gold badge11 silver badge44 bronze badges, papayapapaya 1, gallardo\_diegogallardo\_diego 6911 silver badge11 bronze badge, & KelumKelum 1. (1965, February 1). Tweepy get tweets between two dates. Stack Overflow. Retrieved March 4, 2023, from <https://stackoverflow.com/questions/49731259/tweepy-get-tweets-between-two-dates>
- [2] API. API - tweepy 4.12.1 documentation. (n.d.). Retrieved March 4, 2023, from <https://docs.tweepy.org/en/stable/api.html>
- [3] Contributor, T. T. (2017, November 29). What is support Vector Machine (SVM)? Definition from TechTarget. WhatIs.com. Retrieved March 4, 2023, from <https://www.techtarget.com/whatis/definition/support-vector-machine-SVM>
- [4] Convert Pandas DataFrame to CSV - javatpoint. www.javatpoint.com. (n.d.). Retrieved March 4, 2023, from <https://www.javatpoint.com/convert-pandas-dataframe-to-csv>
- [5] Eisgandar. (2022, December 1).  Twitter sentiment analysis: Hatred speech. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/code/eisgandar/twitter-sentiment-analysis-hatred-speech>
- [6] GeeksforGeeks. (2020, November 12). Seaborn Heatmap - A comprehensive guide. GeeksforGeeks. Retrieved March 4, 2023, from <https://www.geeksforgeeks.org/seaborn-heatmap-a-comprehensive-guide/>
- [7] Grieve, P., & Writer, C. (2022, March 9). Deep Learning vs. machine learning: What's the difference? Zendesk. Retrieved March 4, 2023, from <https://www.zendesk.com/blog/machine-learning-and-deep-learning/>
- [8] JustAnotherArchivist. (n.d.). Justanotherarchivist/snsrape: A social networking service scraper in Python. GitHub. Retrieved March 4, 2023, from <https://github.com/JustAnotherArchivist/snsrape>
- [9] kanncaa1. (2018, July 24). Machine learning tutorial for beginners. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/code/kanncaa1/machine-learning-tutorial-for-beginners>
- [10] karansehal13. (2022, May 5). Twitter sentiment analysis using Naive-Bayes NLP. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/code/karansehal13/twitter-sentiment-analysis-using-naive-bayes-nlp/notebook>
- [11] kazanova, &M. &M. (2017, September 13). Sentiment140 dataset with 1.6 million tweets. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/datasets/kazanova/sentiment140>
- [12] kharde, V., & Sonawane, V. (2016). Sentiment analysis of Twitter data: A survey of techniques. International Journal of Computer Applications, 139(11), 19-24. Retrieved March 4, 2023, from <https://www.ijcaonline.org/research/volume139/number11/kharde-2016-ijca-908625.pdf>
- [13] Leonel, J. (2019, October 9). Supervised learning. Medium. Retrieved March 4, 2023, from <https://medium.com/@jorgesleonel/supervised-learning-c16823b00c13>
- [14] McElwee, K. (2021, April 25). Mistakes to avoid when using Twitter data for the first time. Medium. Retrieved March 4, 2023, from <https://towardsdatascience.com/mistakes-to-avoid-when-using-twitter-data-for-the-first-time-304c3d0ef7a6>
- [15] Meet The Editor Kaitlin Herbert Browse our tech-specific sites or tell us about a new term. Have some feedback? See a definition that needs updating? Let me know!, & Herbert, K. (n.d.). Browse definitions by alphabet. B - B2B to BIT | WhatIs.com - Search Results | {1}. Retrieved March 4, 2023, from <https://www.techtarget.com/whatis/definitions/B>
- [16] Models. Models - tweepy 4.12.1 documentation. (n.d.). Retrieved March 4, 2023, from [https://docs.tweepy.org/en/stable/v2\\_models.html](https://docs.tweepy.org/en/stable/v2_models.html)

- [17] PapaSmurfPapaSmurf 4377 bronze badges. (1965, April 1). For each trend in list, pull the 1000 most recent tweets using Python-Twitter (and remove/exclude retweets). Stack Overflow. Retrieved March 4, 2023, from <https://stackoverflow.com/questions/50420710/for-each-trend-in-list-pull-the-1000-most-recent-tweets-using-python-twitter-a>
- [18] Passionate-Nlp. (2021, August 9). Twitter sentiment analysis. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis>
- [19] Petersson, D. (2021, March 26). What is supervised learning? Enterprise AI. Retrieved March 4, 2023, from <https://www.techtarget.com/searchenterpriseai/definition/supervised-learning>
- [20] ProjectPro. (2022, June 3). Top 50 machine learning projects ideas for beginners in 2023. ProjectPro. Retrieved March 4, 2023, from <https://www.projectpro.io/article/top-10-machine-learning-projects-for-beginners-in-2021/397>
- [21] Rude, B. (2021, August 1). Extracting geographic location information from Twitter. Welcome to my homepage! Retrieved March 4, 2023, from <https://brittarude.github.io/blog/2021/08/01/Location-and-geo-information-in-twitter>
- [22] Sentiment analysis of Twitter data: A survey of techniq - ijcaonline.org. (2016, April). Retrieved March 4, 2023, from <https://www.ijcaonline.org/research/volume139/number11/kharde-2016-ijca-908625.pdf>
- [23] Shah, K. (2022, January 8). 20 machine learning projects that will get you hired in 2022. Medium. Retrieved March 4, 2023, from <https://medium.com/projectpro/20-machine-learning-projects-that-will-get-you-hired-in-2021-a89473f2d2c7>
- [24] The 25 most controversial topics in college. TheBestSchools.org. (2022, September 20). Retrieved March 4, 2023, from <https://thebestschools.org/magazine/controversial-topics-research-starter/>
- [25] Therealsampat. (2021, January 21). Prediction using supervised ML. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/code/therealsampat/prediction-using-supervised-ml>
- [26] Toosi, A. (2019, January 6). Twitter sentiment analysis. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech>
- [27] Twitter. (n.d.). Explore a user's tweets | docs | twitter developer platform. Twitter. Retrieved March 4, 2023, from <https://developer.twitter.com/en/docs/tutorials/explore-a-users-tweets>
- [28] Twitter. (n.d.). Twitter API V2 Tools & Libraries | Docs | Twitter Developer platform. Twitter. Retrieved March 4, 2023, from <https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries/v2>
- [29] Twitter sentiment analysis. Analytics Vidhya. (n.d.). Retrieved March 4, 2023, from <https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/#About>
- [30] vincerlanz09. (2022, August 27). 🗣️ tweets EDA + sentiment analysis. Kaggle. Retrieved March 4, 2023, from <https://www.kaggle.com/code/vincerlanz09/tweets-eda-sentiment-analysis>
- [31] Ysenarath. (n.d.). Ysenarath/tweetkit: A python client for Twitter API. GitHub. Retrieved March 4, 2023, from <https://github.com/ysenarath/tweetkit>