Toxicity and Emotion-Based User Suspension Prediction on Twitter: A Comparative Analysis Study

Nikhil Jindal nikhilgh@usc.edu
Nikhil Kanatala kanatala@usc.edu
Shreya Dwivedi shreyad@usc.edu
Unnati Singhal unnatisi@usc.edu
Yash Bitla bitla@usc.edu

Abstract

This project addresses the critical challenge of enhancing content moderation and ensuring user safety within the digital realm, with a specific focus on Twitter. Our primary objective is the development of a predictive classifier capable of assessing the likelihood of user suspension based on the toxicity and emotion scores associated with their tweets. In addition to this, we will compare various machine learning models to select the most effective model for predicting suspension risk based on these scores. By accurately predicting suspension status, our project supports early intervention, enabling platform administrators to take preemptive measures. Furthermore, we strive to foster a safer and more inclusive online environment by detecting and addressing harmful or offensive content.

1 Project Domain and Goals

This project is centered on the development of a predictive classifier aimed at evaluating the probability of user suspension on social media platforms, with a specific emphasis on Twitter. The fundamental objective is to address the pressing challenge of enhancing content moderation and bolstering user safety within the digital realm. In a world where social media profoundly influences our lives, there is an increasing need for effective mechanisms to identify and mitigate the risks associated with toxic online behavior.

To achieve this goal, we employ Natural Language Processing (NLP) techniques, which are instrumental in quantifying tweet toxicity and emotional underpinnings. We utilize external resources, such as Perspective API (Jigsaw by Google) for toxicity score calculation and the Demux-MEmo tool (Chochlakis) for emotion score analysis, to enrich our project.

Through the successful implementation of our ideas, this project aims to solve several critical problems. Foremost among these is the need for proactive identification of users whose online behavior poses a risk to the digital community. By accurately predicting the suspension status of users based on tweet toxicity and emotion scores, the project contributes to early intervention, allowing platform administrators to take preemptive measures. Additionally, our project seeks to foster a safer and more inclusive online environment by detecting and addressing harmful or offensive content.

2 Related Work

Twitter has been a popular form of media to exchange information about news, entertainment, marketing, and sometimes misinformation about several events. Various studies have shown how misinformation can be propagated through social media. One such paper (Pierri et al., 2023) investigates how information spreads on Twitter and Facebook during the Russian invasion of Ukraine and the creation of new accounts by coordinated user groups boosting spamming and hate speech. Again, (Hanley et al., 2023) study, western, Russian, and Chinese media on Twitter and Weibo discussing topics ranging from political and diplomatic consequences on the geopolitical landscape to humanitarian aspects of war.

Given the above discussion, we need to find a way to stop the spread of misinformation and hate speeches on Twitter. (Salehabadi et al., 2022) started by segregating toxic and non-toxic conversations between Twitter users. (Salehabadi et al., 2022; Saveski et al., 2021) concluded that toxic Twitter threads are generally longer, with only a few users participating. (Saveski et al., 2021) also deduce that toxic replies come from users without social connections or common friends with the poster, resulting in a sparser follow graph. (Chowdhury et al., 2020) suggests analyzing the purged accounts that Twitter has banned to help

analyze the behavior of future toxic accounts used as abuse tools. They check the timeline of the purged accounts to see if they remain dormant, and they suddenly start spreading malicious content. The behavioral pattern in the activity of such accounts is very similar and repetitive that it can be attributed as bot-like (Qayyum et al., 2023) posting coherent and bot-like content using similar hashtags, URLs, and domains.

Overall, our understanding of how to regulate Twitter content and purged users is limited. This paper focuses on integrating toxicity scores and emotional analysis, helping perform predictive tasks to make Twitter and other social media platforms a safer and more respectful online space.

3 Datasets

For this project, we will utilize the Twitter dataset, spanning 4 months (January 1, 2022 - April 24, 2022). This dataset will be compiled through two sources. First, for the period February 22, 2022 - April 24, 2022, we will refer to an existing dataset collected through the Twitter Standard v1.1 Streaming endpoint, which contains tweets matching 30+ conflict-related keywords in English, Russian, and Ukrainian. In May 2022, we will use the Twitter Historical Search API to collect tweets from January 1, 2022 - February 21, 2022, using the same set of keywords. This comprehensive Twitter dataset comprises almost 250 million tweets from 15 million unique users, with an additional column indicating whether each user is ”suspended” or ”active.”

3.1 Data Processing

Data preprocessing will involve cleaning and organizing the raw dataset. Text cleaning, tokenization, and handling of missing or erroneous data will be part of this preprocessing phase. Additionally, we will aggregate data at the user level by calculating toxicity and emotion scores, resulting in a dataset with the number of unique users.

The dataset from this project originates from (Pierri et al., 2023) which provides access to post IDs for Twitter datasets. Data retrieval is done by querying the Twitter API for Twitter data.

This project builds upon the existing dataset, leveraging both raw and preprocessed data to develop a classifier predicting user suspension risk based on toxicity and emotion scores extracted from tweets.

4 Technical Challenges

The primary technical challenge in this project lies in developing an effective prediction model that accurately foretells the suspension risk of Twitter users based on their tweets' toxicity and emotion scores. This challenge encompasses several intricate aspects.

1. **Data Diversity and Quality:** Acquiring a diverse dataset that adequately represents both active and suspended users, while ensuring data quality and reliability, is a foundational challenge.
2. **Data Preprocessing:** Cleaning and preparing raw tweet data for toxicity and emotion score calculation entail intricate processes, including text cleaning, tokenization, and handling missing or erroneous data.
3. **Feature Engineering:** Designing informative features derived from toxicity and emotion scores, particularly for user-level aggregation, presents a significant challenge.
4. **Data Imbalance:** Addressing class imbalance issues in the dataset, where suspended users may be a minority, requires careful consideration. Techniques like oversampling, undersampling, or synthetic data generation may be necessary for balanced model training.
5. **Model Selection:** Choosing suitable machine learning models (e.g., Naive Bayes, Decision Trees, XGBoost) and optimizing their hyperparameters is a non-trivial task.
6. **Evaluation Metrics:** Selecting appropriate evaluation metrics that go beyond accuracy, such as ROC curves, confusion matrices, and precision-recall curves, is crucial for comprehensive model assessment.
7. **Ethical Considerations:** Ensuring ethical compliance and preventing bias in toxicity and emotion scoring tools, as well as the developed classifier, is crucial.

This project goes beyond course content, engaging in real-world data integration, advanced feature engineering, and ethical AI considerations. It applies foundational NLP concepts to address real-world social media challenges, aiming to develop a robust suspension risk prediction classifier.

Individual Contributions

All team members are working closely on the project, with each individual contributing uniformly with respect to other members of the team. Given that, the contributions of members are given in Table 1.

Name	Contribution
Nikhil Jindal	Literature review
	Data Collection and Preparation
	Model Selection and Evaluation
Unnati Singhal	Literature review
	Toxicity Score Analysis
	Feature Engineering
Yash Bitla	Literature review
	Data Collection and Preparation
	Emotion Score Computation
Nikhil Kanatala	Literature review
	Emotion Score Analysis
	Model Selection and Evaluation
Shreya Dwivedi	Literature review
	Toxicity Score Computation
	Feature Engineering

Table 1: Team member contributions

References

Georgios Chochlakis. Demux-memo: Emotion recognition tool.

Farhan Asif Chowdhury, Lawrence Allen, Mohammad Yousuf, and Abdullah Mueen. 2020. On twitter purge: A retrospective analysis of suspended users. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 371–378, New York, NY, USA. Association for Computing Machinery.

Hans W. A. Hanley, Deepak Kumar, and Zakir Durumeric. 2023. "a special operation": A quantitative approach to dissecting and comparing different media ecosystems' coverage of the russo-ukrainian war.

Jigsaw by Google. Perspective api: A machine learning model for detecting toxicity and managing online conversations.

Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2023. Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine. In *Proceedings of the 15th ACM Web Science Conference 2023*. ACM.

Hina Qayyum, Benjamin Zi Hao Zhao, Ian Wood, Muhammad Ikram, Nicolas Kourtellis, and Mohamad Ali Kaafar. 2023. A longitudinal study of the top 1% toxic twitter profiles. In *Proceedings of the 15th ACM Web Science Conference 2023*. ACM.

Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User engagement and the toxicity of tweets.

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA. Association for Computing Machinery.