# Stats 140XP Final Project

Laura Ngo, Suraj Rajan, Vanda Suklar, Ryan Largo, Odeya Russo, Masato Ishizuke

March 4 2023

## Abstract

**Need to fix** This study explores how sentiment analysis can be employed to decode flagged responses within the previous president, Donald Trump's Twitter account both before and during his campaign. We use advanced machine learning techniques to conduct a detailed analysis of the sentiment expressed in such flagged tweets. In doing so, we aim to understand any of Trump's underlying intentions and attitudes to influence his audience. Our findings reveal that models including blank, blank, and blank are better suited for sentiment analysis in high-dimensional spaces. We observed patterns in the sentiment expressed across different periods which shed light on the dynamics of online discourse surrounding political figures. These findings have important implications for understanding social media behavior and its impact on public opinion and political discourse. From this, future research could explore additional factors influencing sentiment analysis outcomes, such as user demographics and linguistic features, to further enhance our understanding of online communication dynamics.

⋆ ⋆ ⋆ ⋆ ⋆

## 1 Introduction

- talk about twitter's flagging policy, for both intro and report - include where the dataset comes from - focus on why flagged tweets predominantly occur during 2020 (is there something with twitter's policy that changed so that election misinformation could be flagged more frequently??)

In the evolving landscape of social media discourse, the intersection between political communication and platform moderation policies has increased in importance. Using a dataset consisting of webscraped tweets from Donald Trump between 2009 and 2021, our study aims to predict through a sentiment analysis which tweets are more likely to be flagged by Twitter's moderation system.

## 2 Preprocessing

•

### 2.1 Data Cleaning Process

For data processing, we initially reviewed Donald Trump's Twitter, separating between his personal tweets and retweets from other users and checking for NA values in our data. We then removed all retweets by converting to boolean value and setting them to false. Following this, we removed stop words and other noise including URLS to reduce the dimensionality of our data and focus on more meaningful terms that contribute to the semantics of the test for classification purposes.

With our cleaned dataset,

- changed to boolean
- took all retweets out and turned to false
- removed stop words such as https
- tokenize, break paragraph into individual words so computer can easier process
- stem words such as happier to happy
- specified x = tweets, y = flagged or not
- flagged and not flagged with count of which are popping up
- more popping up = less importance

- balance with random oversampler, use minority class and consistently random sampled for minority to balance data
- dimension reduction using pca
- transformed text using dimension reduction
- split into training and testing 80,20
- plug into models and testing predictions

### 2.2 Sentiment Lexicon

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## 3 Experiment

We are committed to identifying the most precise and efficient model among a selection that includes Naive Bayes, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression.

### 3.1 Naive Bayes

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| False | 1 | 0.95 | 0.97 | 9356 |
| True | 0.95 | 1 | 0.97 | 9222 |
| Accuracy | | | 0.97 | 18578 |
| Macro Avg | 0.98 | 0.97 | 0.97 | 18578 |
| Weighted Avg | 0.98 | 0.97 | 0.97 | 18578 |

Table 1: Naive Bayes Classification Report

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| False | 1 | 0.97 | 0.99 | 9356 |
| True | 0.97 | 1 | 0.99 | 9222 |
| Accuracy | | | 0.99 | 18578 |
| Macro Avg | 0.99 | 0.99 | 0.99 | 18578 |
| Weighted Avg | 0.99 | 0.99 | 0.99 | 18578 |

Table 2: Random Forest Classification Report

### 3.2 Random Forest

### 3.3 K-Nearest Neighbors

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| False | 1 | 0.97 | 0.98 | 9356 |
| True | 0.97 | 1 | 0.99 | 9222 |
| Accuracy | | | 0.98 | 18578 |
| Macro Avg | 0.99 | 0.99 | 0.98 | 18578 |
| Weighted Avg | 0.99 | 0.98 | 0.98 | 18578 |

Table 3: Random Forest Classification Report

### 3.4 Logistic Regression

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| False | 1 | 0.95 | 0.97 | 9356 |
| True | 0.95 | 1 | 0.97 | 9222 |
| Accuracy | | | 0.97 | 18578 |
| Macro Avg | 0.97 | 0.97 | 0.97 | 18578 |
| Weighted Avg | 0.97 | 0.97 | 0.97 | 18578 |

Table 4: Logistic Regression Classification Report

# 4 Conclusion and Summary

## 4.1 Results and Analysis

| Model | Accuracy (%) Before/After PCA |
|---|---|
| Naive Bayes | 85 %, 84 % |
| Random Forest | 83 %, 84 % |
| KNN | 84 %, 81 % |
| Logistic Regression | 56 %, 79 % |

Table 5: Model Accuracy Comparison

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## 4.2 Final Remarks

Based on our results...

# 5 References