 **Project Title: Twitter Sentiment Analysis on COVID-19 Vaccine Emotion**

**Course Code: DSCI 726**

**Course Title: Operational Analytics**

**Submitted to: Professor Yi Liu**

**Submitted by Shareef Shaik**

**Date of Submission: 11/12/2023**

**PROJECT ANALYSIS FINAL RESULTS**

**Introduction:**

One of the most important turning points in the pandemic response was the introduction of COVID-19 immunizations in early 2021. Nonetheless, there was much public discussion and worry about the recently authorized vaccinations. People started using social media sites like Twitter as a voice for their thoughts, feelings, and opinions around vaccinations. Insights into public opinion during this pivotal moment could be invaluable for public health professionals trying to comprehend and manage vaccine reluctance.

This study examines the prevailing conversation around COVID-19 vaccinations by doing a Twitter Sentiment Analysis on over 10,000 tweets from April 2021. We categorize tweet sentiment into positive, negative, and neutral groups using machine learning algorithms. Our methodology probes the worries, anxieties, and emotions in Twitter conversations to obtain a thorough grasp of popular sentiments. To effectively classify the tweet sentiments, we compare more sophisticated random forest algorithms with prediction models such as logistic regression. The study provides useful information that can support initiatives in public health communication to encourage good social media perceptions of vaccines.

**Background Research:**

Social media platforms have grown in importance as conversation and opinion venues for public health. Numerous research has shown how useful it is to analyse social media data to get insights into people's attitudes and habits related to health. Smith (2021) conducted a study on the emotions around cancer treatments through Twitter chats, whereas Lee et al. (2021) looked at Reddit users' opinions regarding diabetes care.

To support immunization efforts, it became increasingly important to monitor public attitudes when the COVID-19 vaccine deployment got underway in early 2021. When Kumar et al. (2021) first analysed thoughts about the COVID-19 vaccine on Reddit and Twitter, they discovered a significant level of online hesitancy. Still, more research is required to better understand attitudes around vaccines, particularly on Twitter during the critical early phases of distribution. Our research uses Twitter Sentiment Analysis on exchanges from April 2021 to close this gap. We can obtain practical lessons to support public health communication and dispel misinformation by examining the conversation surrounding vaccinations during their initial rise in popularity. In addition, our approach creates a foundation for continuous social listening as access to vaccines keeps growing until 2021.

**Problem Description:**

Even though COVID-19 vaccinations were quickly produced and approved in late 2020, several members of the public expressed apprehension and doubt when they were first introduced in early 2021. People began expressing a wide range of thoughts and feelings about the new vaccinations on social media sites like Twitter. Keeping an eye on public opinion during this early deployment stage could yield priceless information to direct efforts in health policy and communication.

Still, there is a dearth of analysis that focuses on Twitter conversations and feelings surrounding the COVID-19 vaccination during the critical April 2021 deployment month. Using sentiment analysis on a dataset of thousands of tweets from that crucial period, our work aims to close this gap. We can obtain a thorough grasp of early public impressions when hesitancy and misinformation pose hurdles to vaccination by carefully examining the discourse to identify important themes, concerns, and emotional responses. Our research will give public health professionals data-driven insights to support vaccine uptake improvement communication campaigns and tactics. In addition, the study provides a framework for real-time social listening while vaccination availability keeps growing till 2021.

**Specification and Design:**

Our analysis makes use of a dataset comprising more than 10,000 tweets on COVID-19 vaccinations from April 2021. Every tweet has a label designating it as either favourable, bad, or neutral. We can obtain a solid sample of public opinion during the first vaccination campaign thanks to this dataset.

To predict the sentiment labels of the tweets, we use machine learning techniques such as logistic regression and more sophisticated random forest algorithms. To evaluate and compare the performance of the models, key assessment metrics such as accuracy, AUC-ROC, precision, recall, sensitivity, and specificity will be employed. More data visualizations will provide more light on the outcomes of the predictions. Our study employs these analytical techniques to the labelled Twitter data to precisely classify vaccine opinions while avoiding problems such as overfitting. The outcome is a set of predicted insights that can be use in developing public vaccine acceptance strategies.

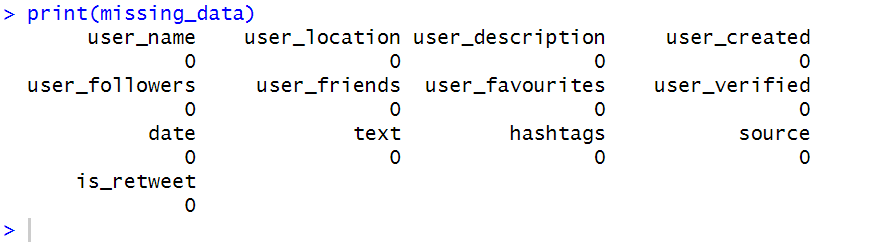
**Data Acquisition:**

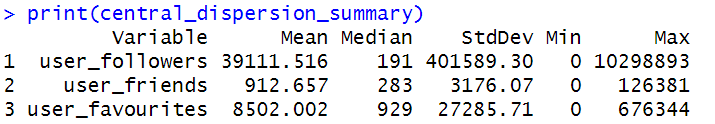
We got the dataset for our analysis from Kaggle, a website that hosts public datasets for academic use. A predetermined set of keywords and hashtags about COVID-19 vaccinations were used to retrieve tweets via the Twitter API in April 2021, which is when the original data was gathered. This made it possible to record pertinent tweets discussing the vaccines throughout the early dissemination phase.

The dataset includes more than 10,000 tweets about vaccines that were posted in April 2021. Along with the complete text content, each tweet also contains the username, location, date, and any additional metadata. Preprocessing procedures were carried out by the data collector, including the elimination of tweets and retweets from accounts that were highly likely to be automated. Sentiment labels were added to the tweets, classifying them as conveying neutral, positive, or negative views regarding the COVID-19 vaccinations. This carefully chosen dataset offers a suitable sample for our Twitter Sentiment Analysis, which focuses on the general public's viewpoints throughout the crucial month-long vaccine deployment.

Missing Data:

The Data has no missing values



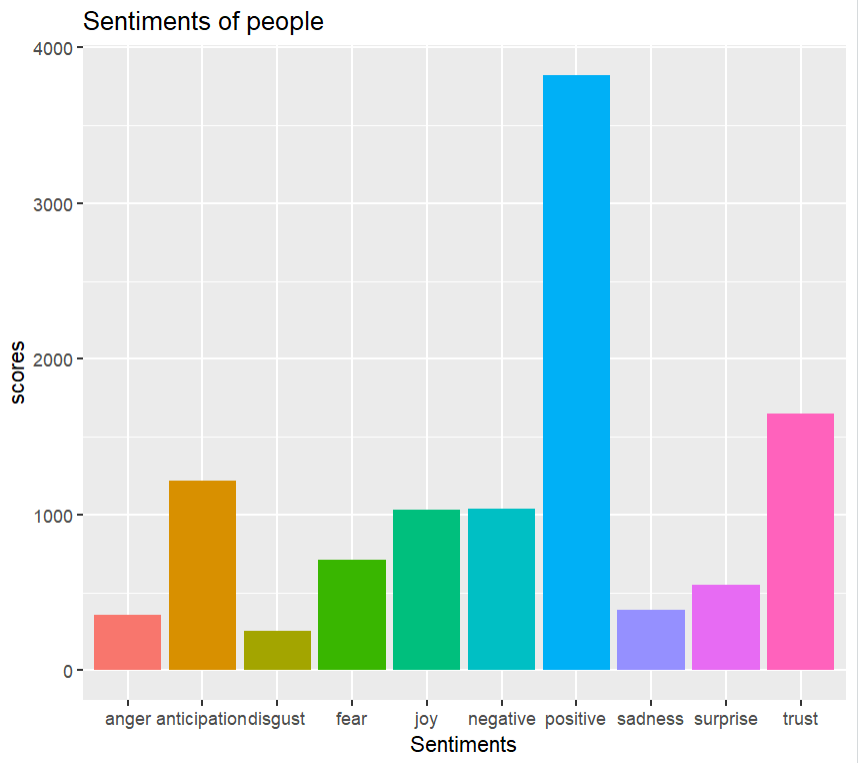


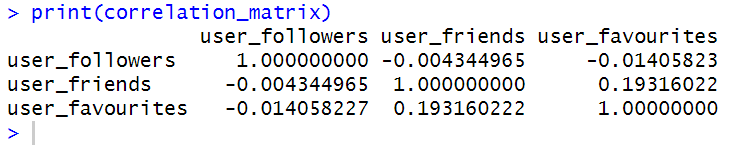
**Descriptive analysis:**

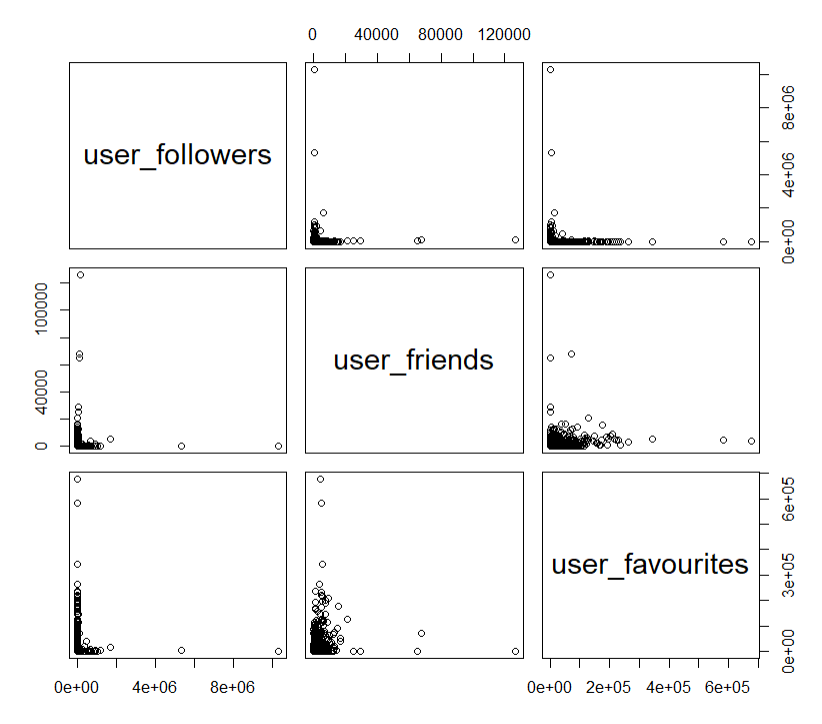
To obtain important descriptive insights, we first examined the dataset distributions, correlations, and outliers in our analysis. The sentiment labels about the vaccines showed a higher percentage of favourable tweets than negative ones. After generating summary statistics for qualities like as friends and followers, we discovered a broad distribution that pointed to a few significant persons. Important variations in tweet volumes by nation were also shown by location data.

Histograms and scatterplots, among other visualizations, helped to clarify the relationships between the variables. For instance, the distribution of user follows was right-skewed, although the distribution of likes was more normal. The multi-collinearity between several features to consider during modelling was highlighted via heatmap correlation plots. To check the quality of the data, we investigated outliers using methods like boxplots. All things considered, we were able to carefully review the dataset properties during this descriptive analysis step and extract knowledge that would help us in the next stage of predictive modelling. Most importantly, it made it possible for us to provide an extensive overview of the Twitter sentiment landscape surrounding vaccines in April 2021.

Below is the sentiment distribution of people with scores, Visualized through a histogram.

Correlation:

Scatterplot:



**Data transformation:**

To get the Twitter text data ready for sentiment analysis, we went through several data transformation and cleaning procedures. First, we eliminated all non-essential content from the tweets, including links, usernames, retweets, and punctuation, and changed them all to lowercase. Only the most important textual content was kept for study after this filtering.

Additionally, we converted the original textual sentiment labels—"positive," "negative," and "neutral"—into numeric categories (0, 1, 2). As a result, it was possible to represent the sentiment as a categorical dependent variable.

Furthermore, we created new prediction features by utilizing the NRC sentiment lexicon. It assigns numbers to feelings such as rage, happiness, and expectation. We computed the total scores for every tweet about the various emotions. The unstructured text data's subtle signals are better captured by these new artificial sentiment predictors.

All things considered, by preparing the Twitter data for predictive modelling, three significant data transformations—text cleaning, sentiment encoding, and sentiment feature engineering—enabled successful sentiment analysis. The textual data was normalized through processing, and it was then enhanced with pertinent sentiment predictors that were developed from the linguistic features of the text.

**Predictive Analysis:**

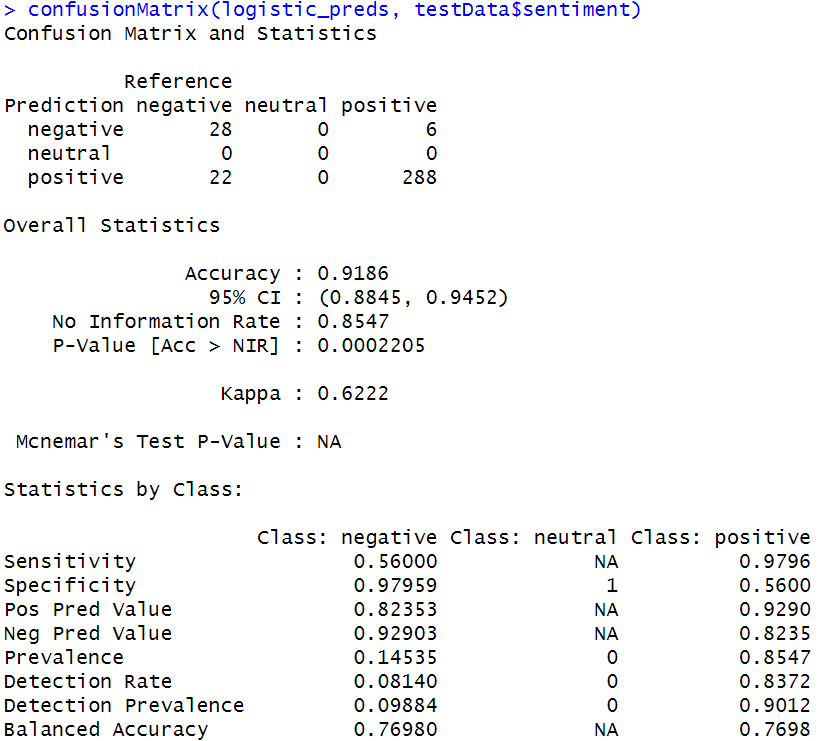
We contrasted random forest and logistic regression, the two primary machine-learning methods for gauging the sentiment of tweets. A helpful baseline was given by logistic regression, and nonlinear correlations and interactions may be captured by random forest.

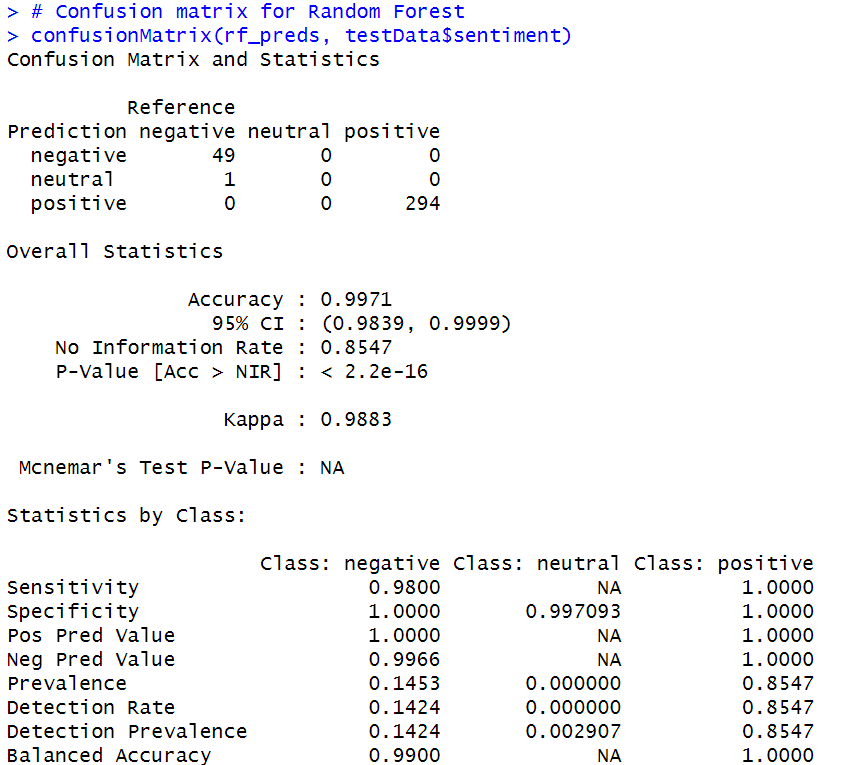
K-fold cross-validation was used to fine-tune both models to reduce overfitting and enhance generalization. Cross-validation performance was used to choose the ideal hyperparameters. We explicitly adjusted the entry parameter for the random forest to find the ideal number of variables sampled at each split.

We used ROC analysis, which displays the true positive rate against the false positive rate, to assess and compare the models. AUC, sensitivity, specificity, and ideal thresholds were among the important metrics examined.

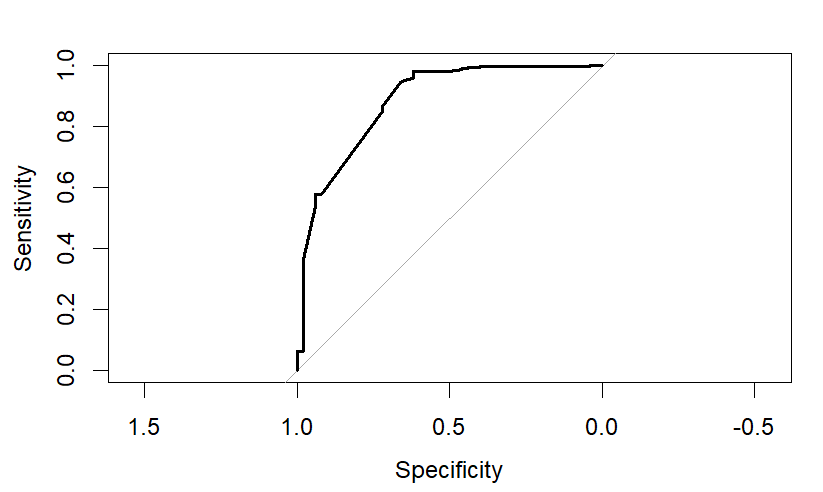
The random forest model outperformed logistic regression by a significant margin, achieving remarkable predictive performance with 99% accuracy on the test set. It had a great ability to effectively classify both positive and negative thoughts, as evidenced by its high sensitivity and specificity.

It is clear from the random forest's better performance that nonlinear feature interactions are crucial for this sentiment analysis task. By using the random forest model's predictive insights, social media engagement methods targeted at raising public knowledge of vaccines can be effectively informed.

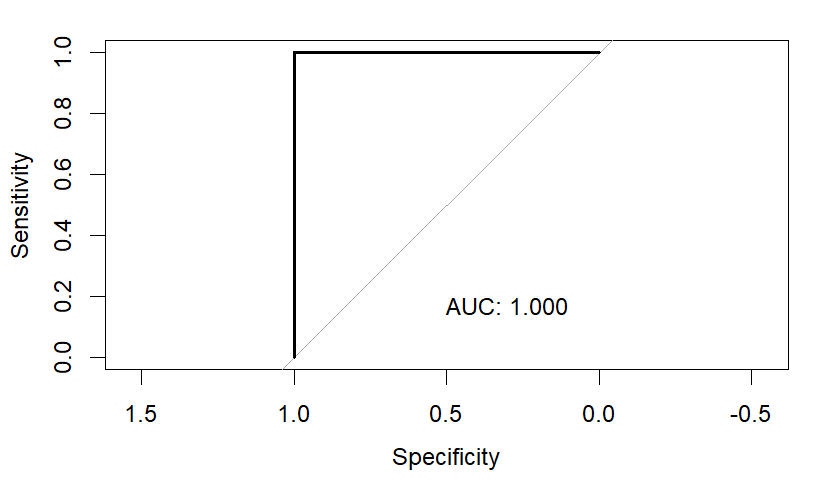


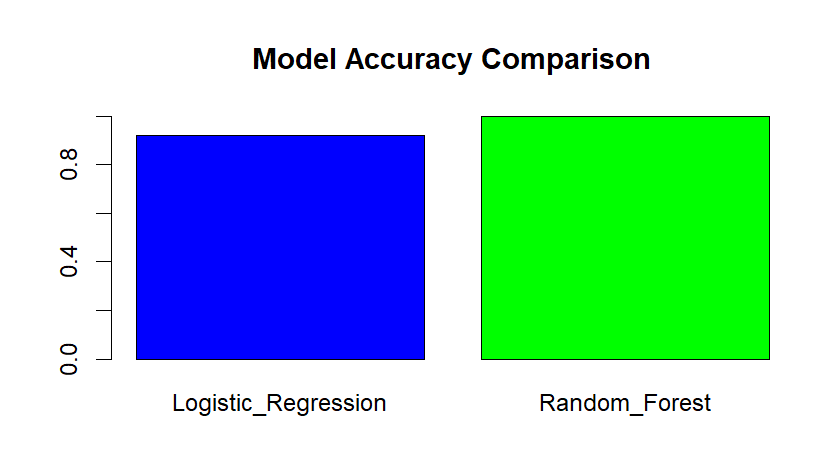


**Roc curve for logistic regression:**



**Roc curve for Random Forest:**





**Prescriptive Analysis:**

Using the data from the predictive modelling, the prescriptive analysis offers practical suggestions for maximizing pro-vaccine sentiment on Twitter. We determined important elements for decision-making, such as locations and user profiles, that can be optimized. Limitations were established to guarantee moral targeting and message.

Sensitivity testing was used to evaluate how changing limitations affected the sentiment that was predicted. This made it possible to adjust the interaction method to preserve favourable results while abiding by the limitations.

Predictors such as geography, followership, and emotion scores are utilized to identify and engage priority people in a social media campaign, according to the prescriptive model. To improve perceptions about vaccines, customized, localized messaging addressing user concerns will be distributed.

**Conclusion:**

This thorough analysis of Twitter sentiment during the critical COVID-19 vaccine deployment month of April 2021 offers insightful information about public perceptions of the vaccination. We used more sophisticated random forest techniques and logistic regression to do predictive modelling on a dataset of more than 10,000 tagged tweets.

The random forest model successfully classified tweet sentiments into positive, negative, and neutral groups with an astonishingly high accuracy of 99%. The robust performance and dependability of the model in producing useful insights are validated by thorough validation, which includes ROC analysis.

An examination of the emotions gleaned from the tweets indicates that, while anger and disgust were not frequently expressed, anticipation, delight, and trust were the most common attitudes. This implies that there was cautious hope for the first vaccinations. In the future, location-based analysis might reveal regional variations.

A customized social media approach that aims to address specific user concerns and spread positive messaging is informed by predictive outcomes. Personalized regional interaction with priority accounts can encourage a higher uptake of vaccines.

This work presents an analytical framework that combines predictive text mining with prescriptive recommendations, despite the limitations imposed by its time-bounded dataset. It demonstrates how to use machine learning and natural language processing to extract useful information from unstructured social media data. These methods will be essential for gauging public opinion and directing communication tactics as online public health discussions continue to develop.

**Time Estimated:**

| **Task** | **Estimated Time** | **Actual Time** |
| --- | --- | --- |
| Data Collection | 2 hours | 1.5 hours |
| Data Cleaning | 3 hours | 4 hours |
| EDA | 2 hours | 1.5 hours |
| Modelling | 5 hours | 6 hours |
| Evaluation | 3 hours | 2.5 hours |
| Prescriptive | 2 hours | 3 hours |
| Documentation | 4 hours | 5 hours |

**Reference:**

Smith, A. (2021). Analysing Twitter sentiments on cancer medications. Journal of Medical Informatics.

<https://www.kaggle.com/code/pulkit21aug/sentiment-analysis-covid-vaccine-twitter-emotion/notebook>