COVID-19 Tweets Sentiment Analysis: An investigation of attitudes and discussions around the COVID-19 pandemic

Ila Kanneboyina, Katherine Lin, Zhengze Zhang, Mateo Alexander, and Hanyi Xu

Quantitative Methods in the Social Sciences, Columbia University

QMSS GR5067: Natural Language Processing for the Social Sciences

Professor Patrick Houlihan

December 16, 2024

Abstract

The COVID-19 pandemic has caused never foreseen levels of discourse online, especially across social media platforms. This research aims to analyze the general sentiments around COVID-19 and key topics of online discourse about the virus. As existing accessible datasets were not continuous, the study focused on four key sub time periods: (1) July to September 2020, (2) December 2020 to February 2021, (3) March to May 2021, and (4) June to September 2021, as well as changes throughout the entire timeline. To investigate sentiment and topics during these time periods, sentiment analysis, word cloud analysis, hashtag frequency analysis, topic modeling, and cluster modeling was performed on existing datasets from X (formerly Twitter) from Kaggle. Results primarily showed that vaccines dominated discourse throughout the COVID-19 dataset. They also generally showed that discourse at the beginning of the time period was more negative, with topics around sickness and death most prominent, but were more positive towards the end as topics shifted more towards distribution of vaccines. Results also showed that negative emotions of anticipation and trust were the highest during the time period. However, these findings are limited because of bots, incomplete tweets, inclusion of irrelevant data (such as birthdays) in analysis, and most of the data set coming from vaccine centered tweets. These findings suggest that the population had a lot of mixed sentiment during the COVID-19 epidemic.

Keywords: COVID-19, pandemic, vaccine, sentiment, sentiment analysis

Introduction

As the world eases out of the COVID-19 pandemic, transitioning back into daily life and reaching some sense of normal, an understanding of public response during the pandemic has become a topic of interest for many researchers. Given the vast amount of data four years after COVID-19 was declared an emergency in the United States and globally, many researchers have sought to conduct research regarding the effects of COVID-19 on individual's long-term health, social media discourse, and more, specifically during the early stages of the pandemic.

A key period during the early stages of the pandemic was between the years of 2020 and 2021, where knowledge and understanding of COVID-19 was limited, and governmental and international organization response was rapidly changing to match the conditions of the pandemic. Different periods of these early stages of the pandemic can be broken down into 5 time periods: July 2020 to September 2020, October 2020 to November 2020, December 2020 to February 2021, March 2021 to May 2021, and June 2021 to September 2021.

Between July 2020 and September 2021, vaccines were still in clinical trials as this was only 3 months after the World Health Organization declared a global pandemic and the United States declared a COVID-19 national emergency (Centers for Disease Control and Prevention). At the time, the death toll reached over 1 million worldwide and became the 3rd leading cause of death in the United States, causing waves of fear across the world about the pandemic (Centers for Disease Control and Prevention).

Between October 2020 and November 2020, the Moderna and Pfizer-BioNTech's vaccines were found to be ~95% effective in clinical trials against COVID-19, and in the

United States, COVID-19 cases surpassed 11 million. Then, between December 2020 and February 2021, the Pfizer and Moderna vaccines were first distributed to the general public in America. At the same time, early variants of COVID-19 began to surface, such as the B.1.1.7, Alpha, Gamma, and Beta variants (Centers for Disease Control and Prevention).

Between March 2021 and May 2021, vaccine distribution continued to different age groups despite the Johnson and Johnson blood clotting cases, and in-person activities were approved to occur with precautions, such as social distancing and masking, for vaccinated individuals (Centers for Disease Control and Prevention).

Lastly, between June and September 2021, the Delta variant of COVID-19 forced a push back to stricter masking guidance as cases surge and concerns are raised about the current vaccine's effectiveness against it. This was also a time period where booster shots for the COVID-19 vaccine were made available for at-risk populations alongside the full FDA approval of the Pfizer vaccine for those older than 18 (Centers for Disease Control and Prevention).

When it comes to understanding social media discourse and attitudes towards the COVID-19 pandemic, existing literature has shown that social media played a crucial role in surveying public attitudes, the spread of false information, assessing mental health, and more. In a study done in 2021 by Tsao et al., they identified how social media can play an essential role in "disseminating health information and tackling infodemics and misinformation" during the COVID-19 pandemic (Tsao et al. 2021). Their review of 81 studies that investigated various topics such as infodemics, public attitudes, mental health, detection or prediction of COVID-19 cases, government responses to the

pandemic, and quality of prevention education videos also revealed that social media platforms, specifically X (formerly Twitter) and YouTube, can be valuable tools for public health agencies and organizations to monitor to manage public knowledge and information about ongoing health crises and events (Tsao et al. 2021).

This study takes findings about the power of social media in understanding various aspects of the pandemic to investigate how attitudes and discussion around COVID-19 changed between the summers of 2020 and 2021 on X. More specifically, this research aims to track how sentiment and emotions towards the COVID-19 pandemic changed and what topics were prominently discussed during different segments of this time period. By building a model to investigate these questions, trends in sentiment and discussion topics in online discourse around the pandemic can be identified, a better understand the power of social media in discussion, discourse, and news can be developed, and identification of how social media can become a useful tool for developing an active understanding of public response to and knowledge of government and international organization responses to global events can occur.

Methodology

Data Collection

In order to conduct a longitudinal analysis of COVID-19 sentiments over time, two datasets were combined to represent data spanning 2020 and 2021. The first dataset, sourced from Kaggle (available here), consisted of 179,108 tweets about COVID-19 from July 2020 to September 2020. These tweets were collected using the Twitter API and a Python script, with a focus on tweets containing the #COVID19 hashtag. Unfortunately, there was no data for the time period between October 2020 and November 2021. The

second dataset, also from Kaggle (available here), consisted of tweets about COVID-19 vaccines, including Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V from December 2020 to November 2021. These were collected using the Tweepy Python package and Twitter API, starting with Pfizer/BioNTech-related tweets and later expanding to include other vaccines. Both datasets included relevant tweet attributes (text, hashtags) and user details (location, followers, favorites).

In order to create the final dataset, the two datasets were concatenated while preserving all columns. Upon merging, it was observed that some tweets were truncated with a link. Recognizing the potential impact on sentiment analysis, the tweets were cleaned using regex to remove any links. However, this remains a limitation of the study, as such tweets were analyzed based on partial text. Another limitation was the presence of “bot” tweets, such as those from automated accounts or gambling advertisements, which do not represent genuine human sentiment. Despite attempts, these bot tweets could not be reliably identified and removed, so they were retained with the understanding that their inclusion might skew the analysis. It was also important to consider that because tweets between December 2020 and November 2021 focused on vaccine related tweets, analysis on that time period would be skewed towards those topics and sentiments around those topics.

Data Preprocessing

The first step of analysis is preprocessing the data to prepare for analysis, more specifically sentiment analysis. Noise, corruption, missing values, and inconsistencies are common problems with raw data that can seriously impair model accuracy and

produce inaccurate predictions. Preprocessing enhances data quality and facilitates knowledge extraction from datasets using techniques including cleaning, integration, transformation, and reduction. Given that machines cannot directly comprehend raw text, video, or photos, it aids in preparing material into a format that machine learning algorithms can efficiently examine. A study performed by Maharana et al. emphasizes that preprocessing has a direct impact on a model's learning ability and is crucial for the generalization performance of supervised machine learning algorithms (Maharana et al., 2022). For this study, lemmatization and tokenization were used to preprocess the data. Both lemmatization and stemming techniques are aimed at finding the basic form of words. The main difference between them is that lemmatization generates a root word (lemma) with actual semantics. It needs to use a corpus such as *WordNet* to find the basic word form, determine the part of speech of the word for accurate lemmatization, and identify stop words through the corpus. Stemming, on the other hand, simply cuts off the end of the word according to the rules, but the resulting stem may not be a real word. For example, for the word "running" lemmatization will provide "run", but stemming may get "runn", which is not a real English word. These differences are the reason why lemmatization was leveraged for preprocessing.

Data Analysis

For primary data analysis, four methods were used: emotion bar charts, word clouds, hashtag analysis, and unsupervised machine learning. For the emotion bar chart, tweet sentiment visualizations were generated based on detected emotions and displayed emotions such as positive, negative, trust, anticipation, fear, sadness, and joy. The frequency of each sentiment was displayed and ranked from highest to lowest to

identify the most prevalent emotions. To build the word cloud, Python's word cloud package was used to generate visualizations to see which words appeared most frequently in tweets and understand their prominence during specific time periods. For sentiment analysis, line charts were generated to display average sentiment scores over time to reveal how people's overall sentiment changes, with scores ranging from 0 to 0.25, where higher scores indicated positive sentiment and lower scores indicated negative sentiment. Lastly, to conduct hashtag analysis, hashtags were collected from “hashtag” columns for analysis. These hashtags were ranked by frequency and the top ten were identified and visualized in a bar chart. These hashtags could cue to which topics were of highest interest in COVID-19 discourse. These analyses were performed for each individual time period and overall. For overall analysis, additional analysis was also performed on emotions using combined emotion radar analysis, heat maps, and trend analysis.

Results by Time Period

July 2020 to September 2020

When analyzing COVID-19 tweets from July to September 2021, trends generally tended to align with the context of the time period – shortly before vaccine rollouts and within three months of the pandemic being declared. The emotions bar graph (Figure 4a) revealed the top five emotions: positive, negative, trust, anticipation, and fear. The word cloud (Figure 1a) offered further insight, showing that fears were predominantly centered around new cases, death, and the spread of Covid-19. Hashtag analysis (Figure 3a) highlighted widespread concern about Covid-19 and lockdowns, with the prominence of the Trump hashtag likely reflecting public reactions to his plans for navigating the

pandemic during a period of heightened fear and anxiety. Finally, the sentiment analysis graph (Figure 2a) showed trends consistent with other analyzed time periods, with one notable distinction: sentiments during this window did not drop below a score of 0.0, a pattern that differed from later periods.

December 2020 to February 2021

When analyzing tweets between December 2020 and February 2021, trends relating to popular discussion topics and sentiment and emotion overtime reflected general positivity during the first vaccine rollouts. The word cloud generated in Figure 1b demonstrates a lot of conversation around the COVID-19 vaccines with specifically “PfizerBioNTech” and “Moderna” standing out alongside other vaccine related language such as “vaccine " and "dose". These results are further supported by a hashtag analysis (figure 3b) that shows that “#PfizerBioNTech” and “#Moderna” were top hashtags for this dataset, with the number of “#PfizerBioNTech” tweets having the highest frequency of 2,946 tweets total. This finding aligns with the context of these months where vaccines were first distributed and key providers would have been popular topics of discussion for the general public. The rest of the hashtags are also mostly related to vaccines, which would be most likely a result of the type of tweets pulled during this time period being more centered around COVID-19 vaccines.

When it comes to sentiment, results from the trend of sentiment analysis (figure 2b) showed fluctuation of sentiment overtime, especially within early January where the sentiment score reached its highest (~0.25) and lowest (~ -0.04) within less than a week. However, analysis on emotional distribution of these tweets (figure 4b) demonstrates overall positive emotions during this time period, with positive emotion tweets having a

frequency of over 5,000 versus other emotions at frequencies less than 1,000 tweets. These results could be explained by vaccine shortages at the beginning of 2021 that were balanced with increased testing and vaccine rollouts for the first vaccine doses.

March 2021 to May 2021

The word cloud (Figure 1c) highlights key themes in COVID-19-related tweets from March to May 2021, focusing on vaccination efforts with terms like "vaccine," "dose," and specific brands such as "Covaxin," "Sputnik," and "Moderna," alongside geographic mentions like "India" and "Canada." The sentiment trend chart (Figure 2c) shows mild positivity overall, with fluctuations, including a dip in April and recovery in mid-May, likely reflecting concerns and favorable developments like vaccine rollouts. The hashtag analysis (Figure 3c) emphasizes vaccine-related terms (#Moderna, #Covaxin) and broader pandemic discussions, illustrating the global and brand-specific focus. The emotion distribution chart (Figure 4c) reveals a dominance of positive emotions such as optimism and trust, with negative tones like fear and anger appearing less frequently, reflecting overall hopeful sentiment tempered by concerns about safety, access, and policy.

June 2021 to September 2021

According to the word cloud (Figure 1d), there was a lot of discussion regarding the first and second doses of the COVID vaccines on the way to becoming fully vaccinated. There was also a lot of discussion about the “Apollo”, which might refer to the COVID-Apollo research project. There were also references to different vaccine providers such as Moderna and Pfizer. Sentiment trends on COVID-19 tweets between June and September 2021 (figure 2d) illustrates fluctuations in public attitudes. The chart

shows significant sentiment dips around early July and mid-August, which might correspond to emerging challenges or negative developments, such as rising case numbers or restrictive policy shifts. Conversely, peaks in positive sentiment suggest optimism, potentially aligned with vaccine rollouts or improved public health metrics. This analysis underscores the evolving emotional and social dynamics of the pandemic as captured through social media discourse.

Analysis of the top 15 hashtags within this time period also highlight public engagement with pandemic-related topics. The most frequently used hashtag, “COVAXIN”, appears followed by “Moderna” and “BBMP”, “Covaxin”, “COVID19”, and “CovidVaccine”, emphasizing the dominant role of vaccine-related topics dominating social media narratives surrounding COVID-19 during this period. Emotion distribution analysis (Figure 4d) also demonstrates that positive emotions dominated the dataset majority, suggesting an optimistic tone during this time. Trust and anticipation indicate positive and forward-looking perspectives, which might be related to pandemic recovery or vaccination efforts. Negative tones including fear, anger, sadness, and discussion appear less frequently during this time period. Emotions like surprise and joy were less common in the dataset. The emotional landscape at this time provides insight into how the public seemed to communicate positively about the pandemic on social media during this timeframe.

Results Overall

The COVID-19 pandemic generated immense public discussion on social media platforms such as Twitter. To analyze sentiment analysis of COVID-19-related tweets across the entire time period, NRCLex and VADER analysis techniques were used to

analyze emotion trends, compare the tools' performance, and explore relationships between user attributes and sentiment scores. While NRCLex identifies nuanced emotions, VADER emphasizes polarity (positive/negative). Divergences between the tools arise from their methodologies, highlighting the need for a combined analysis to fully capture sentiment trends. Therefore, using these two tools together provides more comprehensive insights for this study.

Top Hashtags and Categorization

After cleaning the dataset, hashtags were extracted and grouped into two categories:

(1) Vaccine Manufacturers: “moderna”, “covaxin” , “sputnikv” , “pfizer”, “sinovac”, “sinopharm”, “astrazeneca”, “covishield”, “pfizerbiontech”

(2) General COVID-Related Topics: “covid19”, “coronavirus”, “vaccine”, “covidvaccine”, “covid”, “vaccinated”

As seen in figure 5a, “#covid19” dominated, reflecting global engagement with the pandemic. In addition, vaccine-specific hashtags showed dispersed mentions across brands, while general topics like “vaccine” and “covid19” had higher frequencies, indicating widespread public discourse.

Emotion Scores for Vaccine Manufacturers

From the heatmap and emotion trends (figures 5b and 5c), trust dominated across vaccine manufacturers, particularly for Covaxin, Sputnikv, and Pfizer, reflecting public confidence. Anger and fear were low, but present for brands like Sinovac and Sinopharm, indicating mixed perceptions.

Covaxin showed to have the highest trust (0.076) and notable anticipation (0.055), reflecting optimism and acceptance. Pfizer and Sputnik V also had high trust values (0.061 and 0.070 respectively) but exhibited moderate sadness and fear. On the other hand, Sinovac and Sinopharm had relatively lower trust compared to other brands and had anger scores that were slightly elevated (~0.02), indicating mixed public sentiment. Notably, AstraZeneca registered the lowest joy (0.011) and trust, possibly due to concerns over adverse effects reported during the rollout phase.

Emotion Scores for General COVID-Related Topics

As figures 5d through 5e indicate, in contrast to vaccine manufacturers, general COVID-related topics evoked a broader range of emotions, where anticipation and trust dominated for most topics, but fear and anger were more prominent compared to vaccine brands. Topics like “covid19” and “coronavirus" demonstrated elevated levels of fear (0.060 and 0.061) and trust.

Final results show that “COVID19” and “COVID” had high trust scores alongside elevated fear scores, highlighting the dual narrative of trust in public measures and anxiety over the pandemic. They also showed that “COVIDVaccine” had lower trust and higher anticipation scores, showing hope for vaccine progress but lingering uncertainty and that “vaccinated” had strong trust and anticipation scores, reflecting positive sentiment tied to vaccination campaigns.

Combined Emotion Radar Analysis

Results from a combined emotion radar analysis, as shown in figure 5f, show that for the vaccine manufacturers, trust and anticipation scores dominate and that negative emotions (fear, sadness, and disgust) are relatively subdued, suggesting confidence in

vaccine brands. In contrast, general COVID-related topics had higher scores for fear and anger, showing public distress and uncertainty regarding the pandemic's impacts.

Model Analysis

Comparison of NRC and VADER Sentiment Scores With OLS

Table 1: *OLS Regression of NRC and VADER Sentiment Scores*

Model 1: vader positive ~ NRC positive

OLS Regression Results

Dep. Variable:	vader positive	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	7.926
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	0.00488
Time:	03:30:36	Log-Likelihood:	37563.
No. Observations:	45484	AIC:	-7.512e+04
Df Residuals:	45482	BIC:	-7.511e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1001	0.001	91.002	0.000	0.098	0.102
NRC positive	-0.0117	0.004	-2.815	0.005	-0.020	-0.004

Omnibus:	5553.259	Durbin-Watson:	1.912
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7814.028
Skew:	0.971	Prob(JB):	0.00
Kurtosis:	3.592	Cond. No.	8.86

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Model 2: vader negative ~ NRC negative

OLS Regression Results

Dep. Variable:	vader negative	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	136.7
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	1.52e-31
Time:	03:30:36	Log-Likelihood:	39436.
No. Observations:	45484	AIC:	-7.887e+04
Df Residuals:	45482	BIC:	-7.885e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0859	0.001	76.585	0.000	0.084	0.088
NRC negative	0.0544	0.005	11.694	0.000	0.045	0.064

Omnibus:	4851.066	Durbin-Watson:	1.895
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6548.903
Skew:	0.895	Prob(JB):	0.00
Kurtosis:	3.504	Cond. No.	10.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

To evaluate relationships between NRCLex and VADER scores, the OLS method was used, as shown in table 1. It was expected to see some kind of relationship between the NRC and VADER positive and negative scores, but both models have very low R^2 values (0.000 and 0.003), indicating a lack of strong linear relationships between NRClex

and VADER scores. All regression coefficients are statistically significant ($P < 0.05$), which confirms that the scoring trends of NRClex and VADER are inconsistent for the same set of tweets.

Analysis of User Attributes and VADER Sentiment Scores

To determine whether or not other variables – user attributes –could have correlation to the sentiment scores of the tweets, creating confounds, it was necessary to analyze the relationship between user attributes including user follower, user friends and user favourites, with respect to the average vader compound score. As figures 5g through 5h show, correlations between user followers, user friends, favourites, and VADER compound scores are extremely weak, indicating there is no strong linear relationship between user attributes and sentiment scores.

Correlation Matrix and Linear Regression Analysis

To further investigate this relationship, a simple linear regression for the user attributes with respect to the average vader compound score was performed, resulting in issues with the current model and analysis. The scatter plots, as shown in figures 5i through 5k, reveal that extreme values (e.g., millions of followers, friends, or favourites) dominate the distribution. These extreme values or outliers heavily influence the linear regression line, leading to poor model fit. With 360,000+ tweets, noise in the data further weakens meaningful trends.

To counter these issues, a logarithmic transformation and quantile binning was used to preprocess the data before analysis. Since the extreme values in user attributes create skewed distributions, where a small subset of users distorts the analysis, log transformations can be used to compress the range of extreme values while preserving

relative differences. This reduces the influence of large outliers and brings the data closer to a normal distribution.

Log Transformation and Quantile Analysis

After a log transformation, user attributes were divided into four quantile-based groups from low to very high, making that each group has roughly equal observations, reducing outlier bias. This also allows for clearer trends to emerge by analyzing averages across balanced subsets. The quantile binning enables balanced comparisons across groups, revealing trends masked by outliers.

Figures 5l through 5n demonstrate that when it comes to user followers, lower-follower users post more positive tweets and higher-follower users lean towards neutral or slightly negative sentiment tweets, reflecting their broader and more restrained communication style. These figures also demonstrate that sentiment remains relatively stable for the number of user friends a user has with slight dips for the high number of friends groups followed by a rebound for very high number of friends groups. Lastly, these figures demonstrate that as user favorites increase, sentiment scores decrease, which indicates that highly liked tweets often deal with emotionally charged or serious content.

Overall, these analyses demonstrate how social influence metrics (followers, friends, favourites) have weak but consistent relationships with sentiment polarity, and that high levels of social engagement (e.g., many followers or likes) tend to correlate with neutral or less positive sentiment.

OLS Regression: User Attributes and VADER Compound Scores

Since the correlation value between user attributes and VADER Compound Scores are very low, OLS was used to investigate this relationship further. user attributes and VADER Compound Scores were checked in an OLS regression directly, then another analysis was made after taking log with user attributes.

Table 2 and 3: *OLS Regression Before and After Log Transformation*

OLS Regression Results for user_favourites ~ compound

OLS Regression Results

Dep. Variable:	compound	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	347.4
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	1.68e-77
Time:	03:58:07	Log-Likelihood:	-1.6629e+05
No. Observations:	364802	AIC:	3.326e+05
Df Residuals:	364800	BIC:	3.326e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0864	0.001	130.350	0.000	0.085	0.088
user_favourites	-2.819e-07	1.51e-08	-18.639	0.000	-3.12e-07	-2.52e-07

Omnibus:	771.325	Durbin-Watson:	1.890
Prob(Omnibus):	0.000	Jarque-Bera (JB):	674.060
Skew:	-0.057	Prob(JB):	4.26e-147
Kurtosis:	2.823	Cond. No.	4.60e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.6e+04. This might indicate that there are strong multicollinearity or other numerical problems.

OLS Regression Results

Dep. Variable:	compound	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	391.6
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	5.36e-254
Time:	04:08:14	Log-Likelihood:	-1.6587e+05
No. Observations:	364802	AIC:	3.318e+05
Df Residuals:	364798	BIC:	3.318e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1275	0.002	67.503	0.000	0.124	0.131
log_user_followers	-0.0057	0.000	-22.685	0.000	-0.006	-0.005
log_user_friends	0.0068	0.000	17.604	0.000	0.006	0.008
log_user_favourites	-0.0068	0.000	-23.316	0.000	-0.007	-0.006

Omnibus:	816.082	Durbin-Watson:	1.893
Prob(Omnibus):	0.000	Jarque-Bera (JB):	722.915
Skew:	-0.066	Prob(JB):	1.05e-157
Kurtosis:	2.826	Cond. No.	34.7

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

After a log transformation, the R-squared became 0.003 (table 3), which is a slight improvement to the R-squared value of 0.001 before the log transformation (table 2), but it is still quite low. The log transformation slightly improved the model's explanatory power but it remains very weak. In addition, user attributes (followers, friends, favourites) have a statistically significant but practically negligible relationship with the compound. The low R-squared value indicates that user attributes are not strong predictors of sentiment and other factors (e.g., tweet content) likely play a more significant role.

VADER and NRCLex Models **vs Transformer Models**

After analyzing sentiment scores using VADER and NRCLex models, an unsupervised machine learning approach was implemented for comparative purposes.

Transformer models, such as DistilBERT, are deep learning architectures based on the self-attention mechanism, which enables them to capture long-range dependencies and contextual relationships within text. This choice was motivated by the limitations of traditional word vectors such as Word2Vec, which assign fixed representations to words. By contrast, Transformer-based models like DistilBERT generate dynamic embeddings that account for the context in which words appear, allowing for more nuanced sentiment analysis.

The analysis aimed to perform sentiment analysis on pandemic-related tweets by leveraging DistilBERT for extracting word embeddings. Each word is represented as a 768-dimensional vector, capturing its semantic and contextual meaning within a sentence. To simplify the high-dimensional embeddings for interpretation, Principal Component Analysis (PCA) was applied to reduce the dimensions to 2D. The first principal component (PC1) is particularly emphasized, as it captures the largest variance in the data and serves as a proxy for sentiment direction. This approach enabled the mapping of words to a lower-dimensional space while preserving meaningful relationships between them.

The resulting PC1 and PC2 values, along with their corresponding words, were saved into a dictionary. This process effectively created a new sentiment analysis tool analogous to rule-based models such as VADER or NRCLex, where PC1 serves as a dynamic sentiment score. To evaluate its performance, sentiment scores derived from PC1 were compared with VADER compound scores, a traditional sentiment analysis method. Visualizations were generated to explore relationships between PC1-based scores and VADER scores, as well as to analyze trends in word sentiment and

frequency. This comprehensive process demonstrates a novel method for unsupervised sentiment analysis using Transformer-based contextual embeddings.

Analysis and Visualization of the Transformer Model

Table 4: *Average PC1 Scores*

	All Words	Stop Words	Positive Words (Hu and Bing 2004)	Negative Words (Hu and Bing 2004)
Average PC1 Score	-5.86e-06	-0.562	0.012	0.011

Using PC1 values, average sentiment scores were tested across various word groups. The overall mean sentiment score for all words was approximately -0.00, suggesting a near-neutral average. For stopwords, which were expected to have no sentiment polarity, the mean score was 0.56, indicating a deviation from neutrality. Positive and negative word lists sourced from Hu and Bing (2004) were analyzed next. The average score for positive words was 0.01, while the score for negative words was also 0.01. This highlights that while PC1 shows slight polarity for pre-defined sentiment words, it does not strictly adhere to the expected directional bias.

This figure with a bar chart of PC1 values for frequent words and a scatter plot comparing PC1-based scores with VADER compound scores, further illustrate these patterns. The bar chart reveals that high-frequency words such as "vaccine" and "dose" exhibit lower PC1 scores, indicative of negative sentiment associations during the pandemic context. Conversely, words like "for" and "are" display positive PC1 values, despite their lack of inherent sentiment.

As fig 6b illustrates, the weak correlation (Pearson coefficient ~0.036) between PC1 scores and VADER compound scores highlights a fundamental distinction in their

sentiment capture: VADER's rule-based approach assigns fixed scores to predefined lexicons, whereas PC1 scores, derived from unsupervised learning, reflect sentiment dynamically based on contextual relationships within the data. This suggests that PC1 captures more nuanced, context-driven sentiment patterns that traditional lexicon-based methods may overlook.

Table 5: *OLS Regression on VADER Compound Scores and PC1 Compound Scores*

OLS Regression Results

Dep. Variable:	new_compound	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	222.2
Date:	Sun, 15 Dec 2024	Prob (F-statistic):	3.11e-50
Time:	23:31:47	Log-Likelihood:	-2.4310e+05
No. Observations:	364802	AIC:	4.862e+05
Df Residuals:	364800	BIC:	4.862e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1340	0.001	-167.840	0.000	-0.136	-0.132
compound	0.0304	0.002	14.906	0.000	0.026	0.034

Omnibus:	66655.320	Durbin-Watson:	1.840
Prob(Omnibus):	0.000	Jarque-Bera (JB):	746899.545
Skew:	0.551	Prob(JB):	0.00
Kurtosis:	9.923	Cond. No.	2.64

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The OLS regression results indicate that the relationship between VADER compound scores and new_compound (derived from PC1) is extremely weak. While the coefficient of compound (0.0304) is statistically significant ($p < 0.05$), the R^2 value is a mere 0.001, suggesting that the VADER compound score explains almost none of the variation in the new_compound.

This result reinforces the earlier observation that VADER and PC1 capture sentiment through fundamentally different mechanisms. VADER relies on predefined

word lexicons and fixed rules, which assign scores to individual words. In contrast, new_compound values derived from PC1 reflect sentiment through a data-driven, context-aware representation learned by the Transformer model (DistilBERT). The extremely low explanatory power suggests that the two methods, while both designed for sentiment analysis, operate on distinct conceptualizations of word meaning and sentiment, with minimal overlap in their outputs.

Results for the Transformer Model

The results suggest that PC1 effectively distinguishes sentiment to some extent, particularly for words associated with strong contextual meanings. However, deviations such as stopwords showing non-neutral scores highlight limitations in using an unsupervised PCA-based approach alone for precise sentiment detection. The PCA method captures variance in the embedding space, but this variance does not always directly align with sentiment polarity as defined by human interpretation or rule-based models like VADER.

Moreover, the analysis shows that the relationship between PC1 scores and VADER scores is tenuous. This can be attributed to the dynamic nature of Transformer-based embeddings, where word meaning is influenced by surrounding context, unlike VADER, which uses fixed lexicon-based scores.

By applying DistilBERT embeddings and PCA, an alternative sentiment analysis method was developed. This approach leverages context-aware word embeddings to derive sentiment scores, similar to VADER and NRCLex tools. The analysis of the model revealed that PC1 captures some degree of sentiment polarity but deviates for certain word groups like stopwords and that positive and negative words exhibit only marginal

differences in PC1 scores, suggesting subtle sentiment differentiation. Analysis also showed that comparisons with VADER show weak correlation, highlighting fundamental differences in the methodologies.

While the PCA-based approach adds a valuable perspective to sentiment analysis, its results underscore the complexity of modeling sentiment in natural language and the need for hybrid approaches to achieve more robust outcomes.

Discussion

Sentiment analysis revealed that vaccine manufacturers elicited higher trust, anticipation, and positive sentiment compared to general COVID-related topics, which showed more emotional diversity, specifically in elevated fear and sadness. More specifically, PfizerBioNTech, Sinopharm, and Covaxin emerged as the vaccine manufacturers with the highest positive sentiment scores. Results also showed that trust and anticipation dominated emotional responses, reflecting public optimism and confidence in vaccines as a solution to the pandemic. In contrast, general COVID-related topics such as “covid19” and “coronavirus” exhibited higher levels of fear and sadness, highlighting the ongoing public distress, uncertainty, and frustration surrounding the pandemic. It is important to note that the current dataset focuses more heavily on vaccine related tweets, which could have impacted the results from analysis on the data regarding sentiment and topic analysis for COVID-related topics alongside the presence of bot data, incomplete text per tweet, and irrelevant data included in these analyses.

It was also found that the weak correlations observed between NRCLex and VADER scores underscore their methodological differences. NRCLex, which focuses on word-level emotional intensity, often diverges from VADER, which evaluates

sentence-level sentiment polarity. This discrepancy emphasizes the importance of combining tools to gain a more comprehensive understanding of public sentiment. Future studies can explore methods to reconcile these differences, such as integrating both tools with machine learning models.

Regarding potential confounding variables and data points, extreme values in attributes, such as users with an unusually high number of followers or likes, further complicated the analysis. However, techniques like log transformation and quantile binning helped mitigate outlier effects, but the overall relationships remained weak. Furthermore, analysis of user-level metadata (followers, friends, favourites) demonstrated that these attributes had a negligible influence on sentiment scores. While statistically significant relationships were observed after log transformation, the explanatory power remained extremely low ($R^2 < 0.003$) which suggests that user-level metadata alone cannot adequately explain sentiment variations. Instead, the content of the tweets (e.g., text, hashtags, and temporal factors) likely plays a far more significant role, suggesting that future studies should prioritize text-based features, such as word embeddings, topic modeling, and temporal analysis, to better understand public attitudes.

Conclusion

The COVID-19 pandemic sparked unprecedented levels of discourse online, offering researchers a unique opportunity to explore sentiment during the COVID-19 pandemic. This study leveraged various natural language processing techniques such as sentiment analysis, hashtag frequency analysis, word clouds, cluster modeling, transformers, and other machine learning techniques to examine how feelings, emotions,

and attitude evolved between July 2020 and September 2021 based on a concatenated dataset from X (formerly Twitter) available on Kaggle. Many conversations referenced geographic regions like “country”, “India”, “asia”, alongside the hashtags of #Bengaluru, #Sinovac, and #Pfizer, which illustrates the intersection of global and local discussions that include countries like India in Asia. There were also mentions of different COVID variants such the “delta” variant and mentions of prevention and treatment methods such as “hospital” and “clinics”.

Preliminary findings showed that the resources were dominated by fear and uncertainty, especially surrounding sickness and death at the beginning of the syndemic, whereas later public emotion seemed to shift to be more positive with the vaccine rollouts. Furthermore, this study revealed how VADER, NRCLex natural language, and BERT libraries can be used collectively for comprehensive emotional analysis. Despite developments in a more comprehensive analysis, the data posed a large limitation due to the potential of data incompleteness and contextually irrelevant tweets in skewing results.

In the future, researchers can determine how to build on these findings by integrating more advanced language learning models, such as popular transform-based models called large language models (LLMs). Researchers should also consider exploring the interrelationship between public sentiment, policy decisions, and urban planning decisions among health crises and emergency management. Additionally, examining sentiment in other major pandemics and disease epidemics could prove crucial to helping experts better understand and implement crisis communication strategies. Overall, this study provides a deep contribution to how digital communication

platforms like X (formerly Twitter) give voice to citizens to express how they feel, and how these online dialogues guide and shape the public course during pandemics, ultimately fostering a foundation for both academic and practical applications such as crisis management in public health, public administration, and urban planning.

References

Centers for Disease Control and Prevention. (2023). *CDC Museum Covid-19 Timeline*. Centers for Disease Control and Prevention. https://www.cdc.gov/museum/timeline/covid19.html#

Gabriel Preda. (2021). *COVID-19 All Vaccines Tweets [Data set]*. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2845240

Gabriel Preda. (2020). *COVID19 Tweets [Data set]*. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/1451513

Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Retrieved from https://pypi.org/project/vaderSentiment/

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3, 91-99.

Tsao, S.F. et al. (2021). What social media told us in the time of COVID-19: a scoping review. *The Lancet: Digital health, 3(3),* 175-194. https://doi.org/10.1016/S2589-7500(20)30315-0

Appendix

Figure 1: Word Cloud of Different Tweet Concepts

Figure 1a. Jul 2020 - Sep 2020

Figure 1b. Dec 2020 - Feb 2021

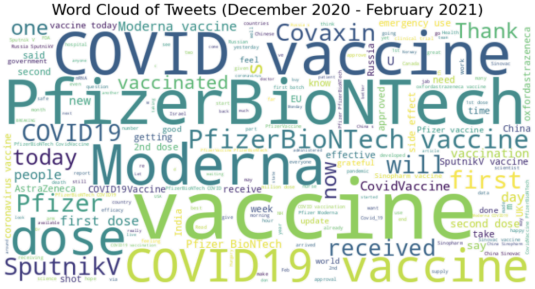


Figure 1c. Mar 2021 - May 2021

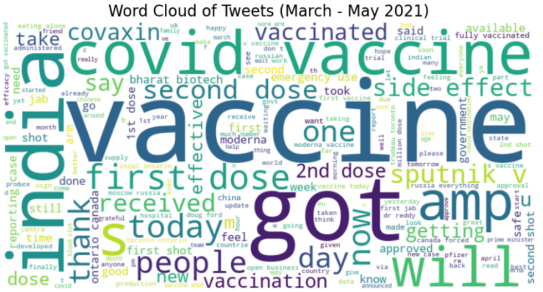


Figure 1d. Jun 2021 - Sep 2021

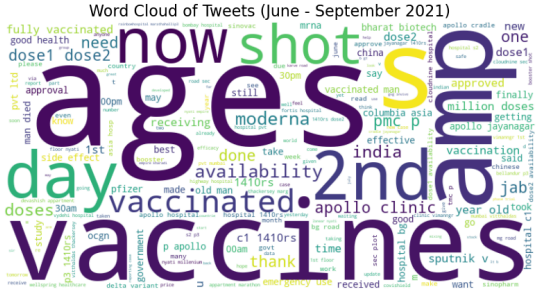


Figure 2: Sentiment Trend Over Time

Figure 2a. Jul 2020 - Sep 2020

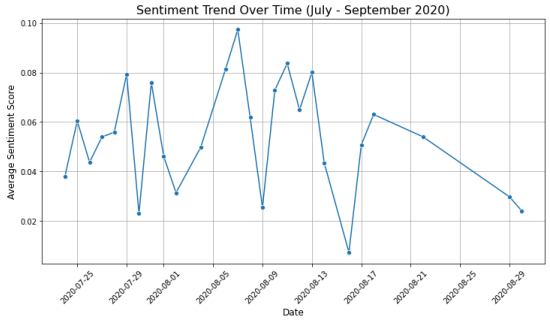


Figure 2b. Dec 2020 - Feb 2021

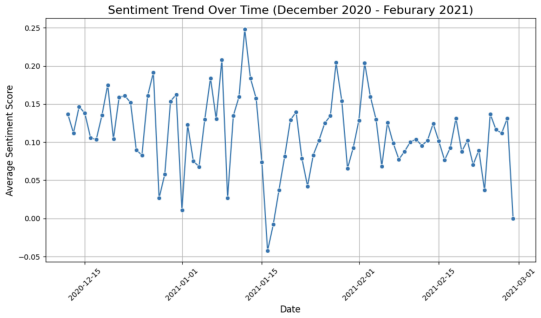


Figure 2c. Mar 2021 - May 2021

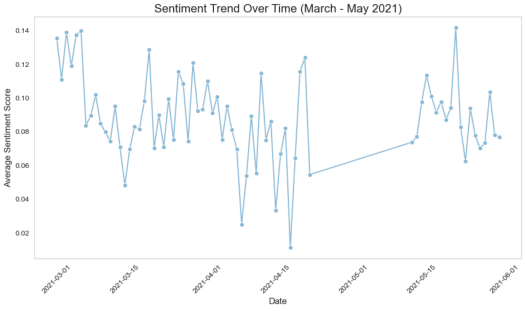


Figure 2d. Jun 2021 - Sep 2021

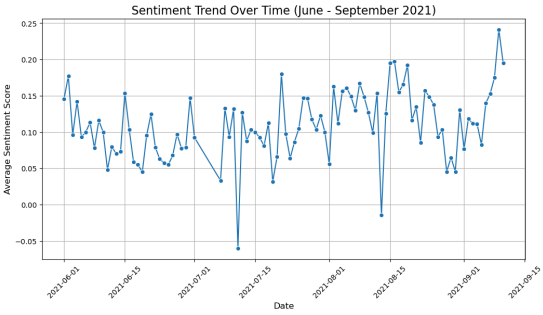


Figure 3: Hashtag Analysis

Figure 3a. Jul 2020 - Sep 2020

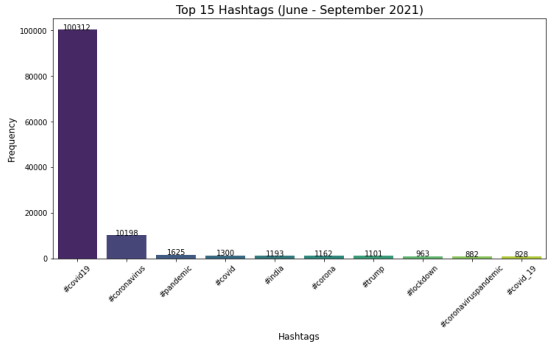


Figure 3b. Dec 2020 - Feb 2021

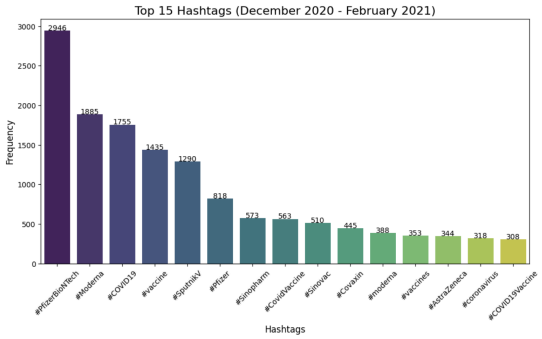


Figure 3c. Mar 2021 - May 2021

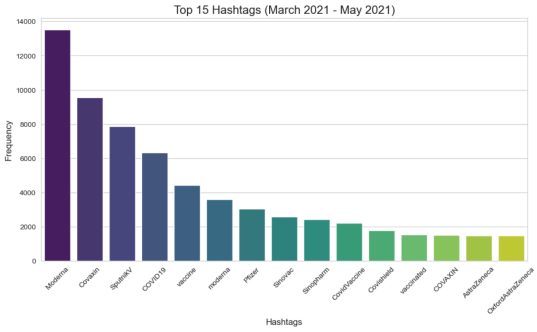


Figure 3d. Jun 2021 - Sep 2021

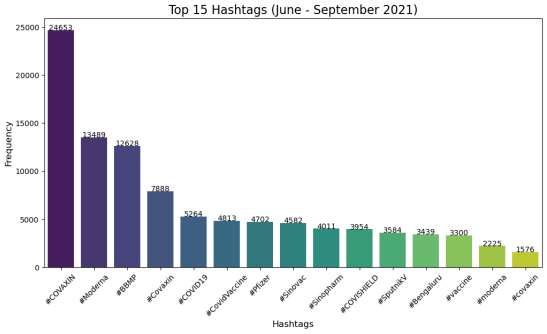


Figure 4: Emotion Distribution of Tweets

Figure 4a. Jul 2020 - Sep 2020

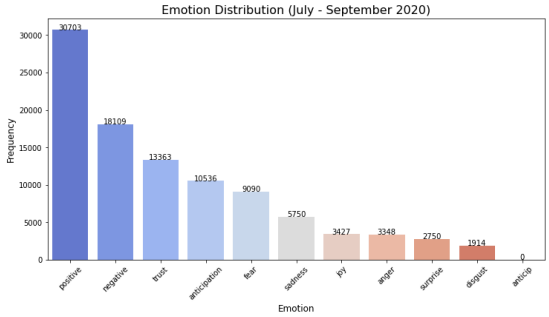


Figure 4b. Dec 2020 - Feb 2021

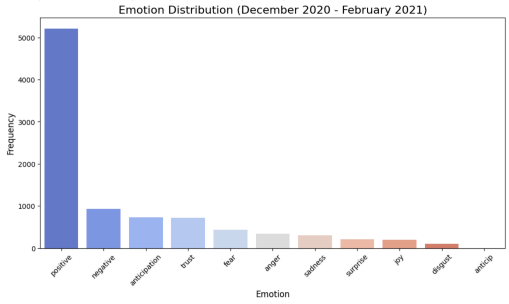


Figure 4c. Mar 2021 - May 2021

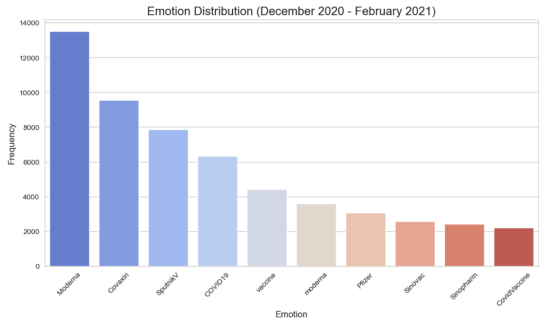


Figure 4d. Jun 2021 - Sep 2021

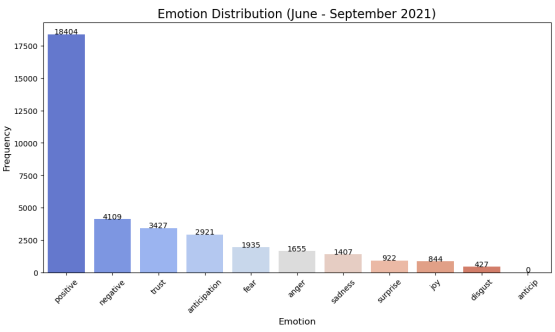
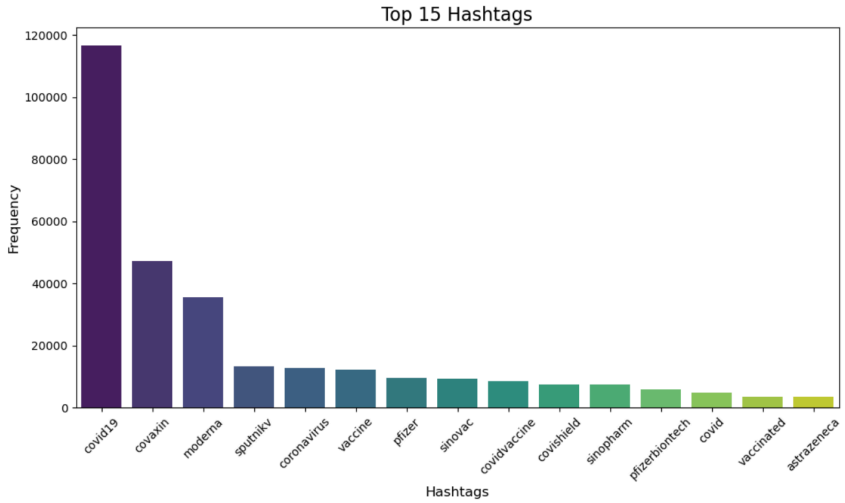
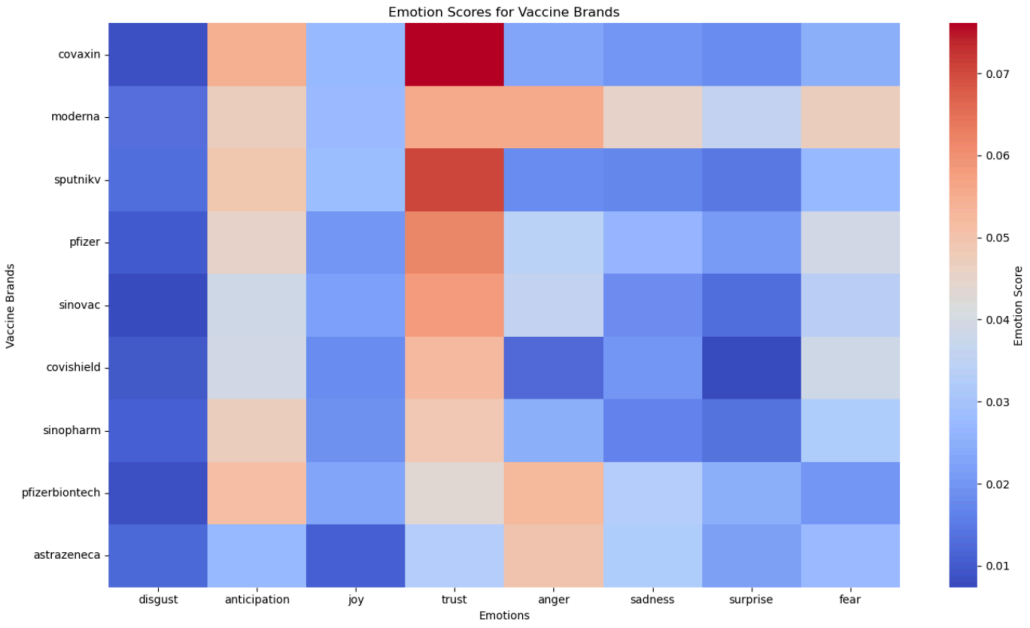
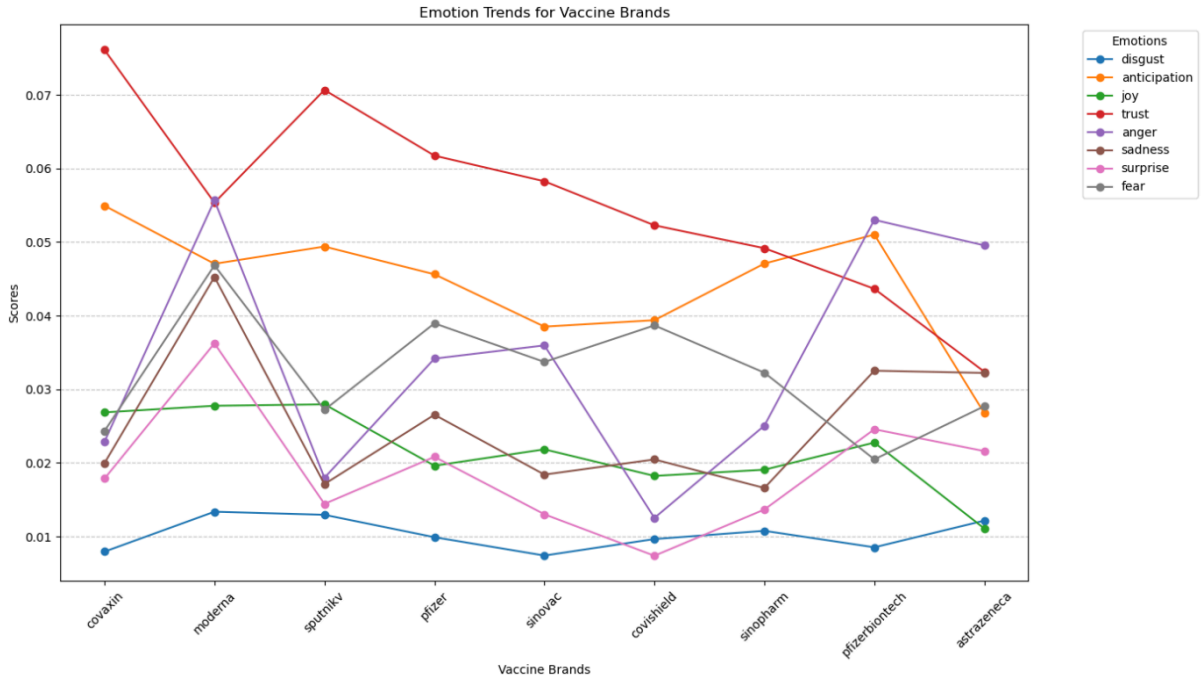


Figure 5a. Top Hashtags and Categorization

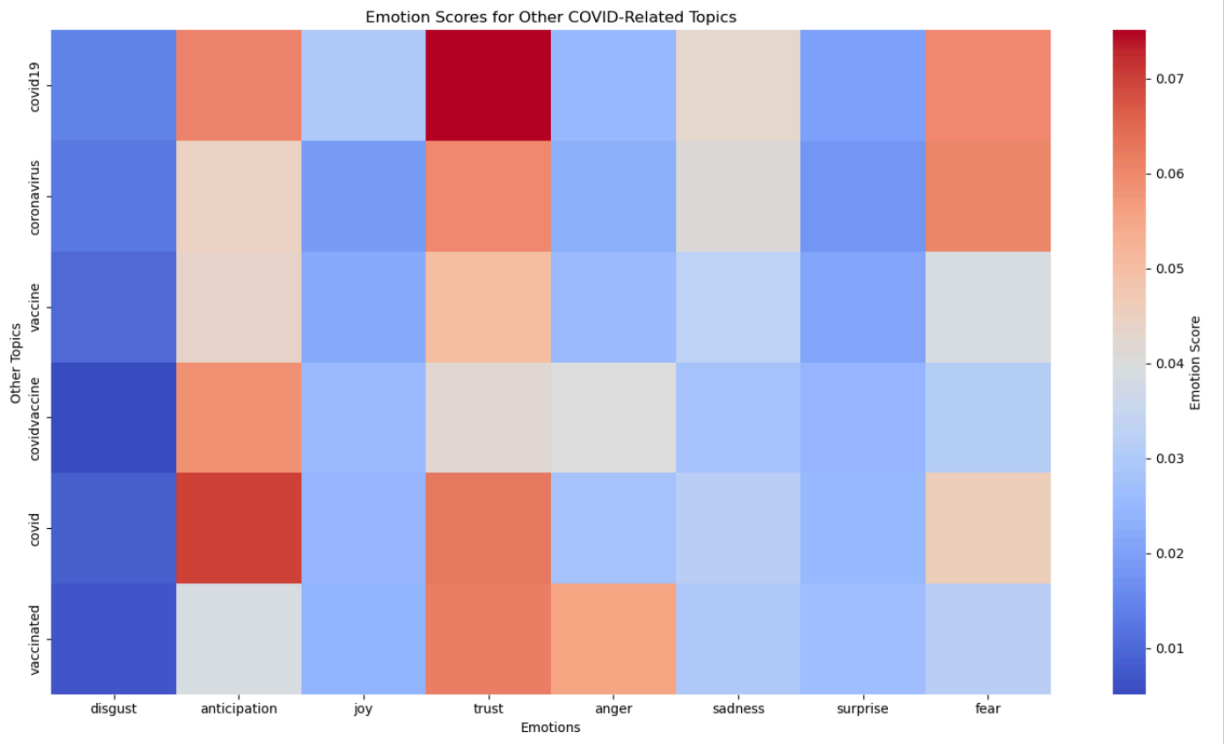


Figures 5b-5c. Emotion Scores for Vaccine Manufacturers Heatmap and Emotion Trends





Figures 5c-5d. Emotion Scores for General COVID-Related Topics Heatmap and Emotion Trends



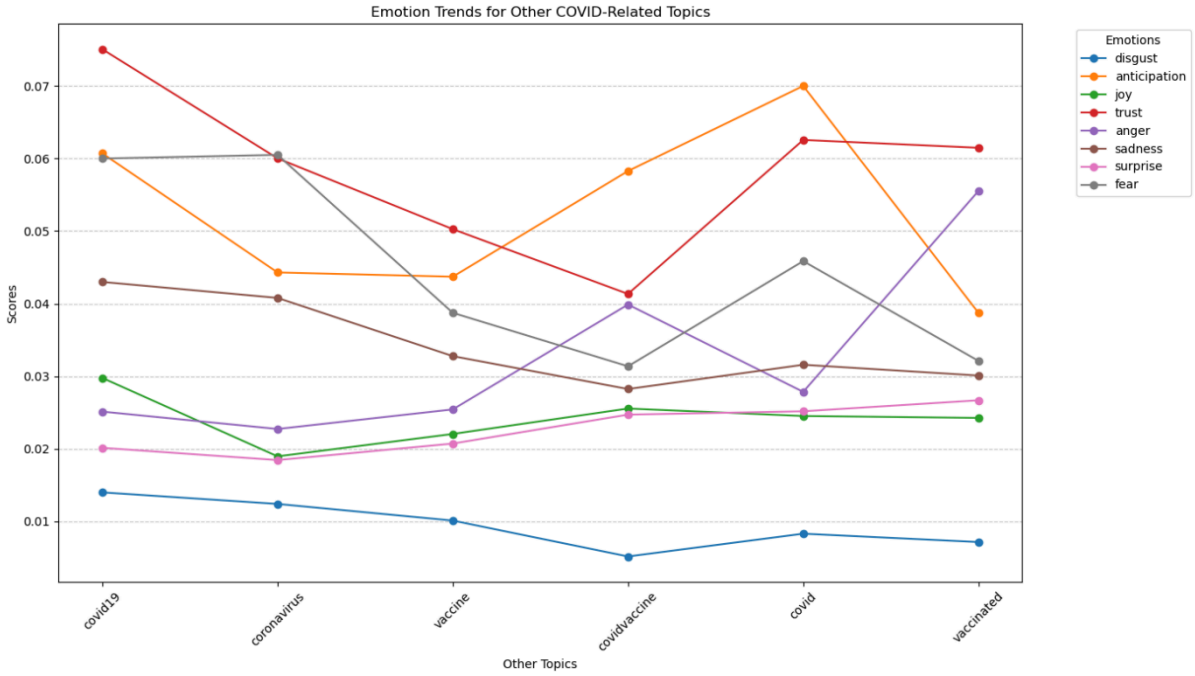
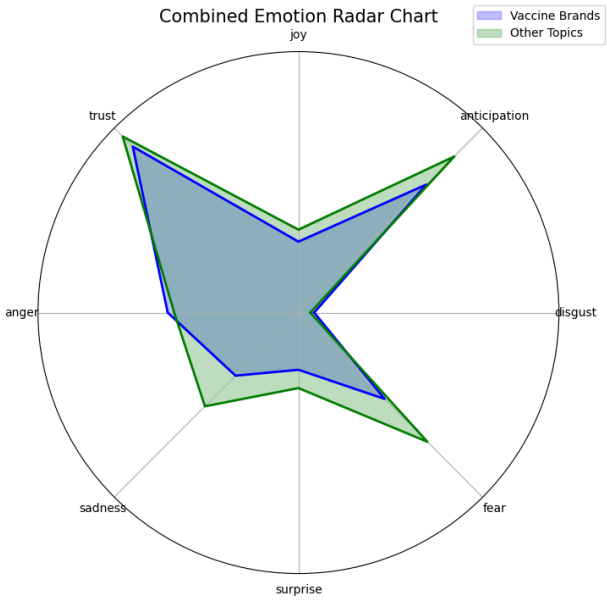
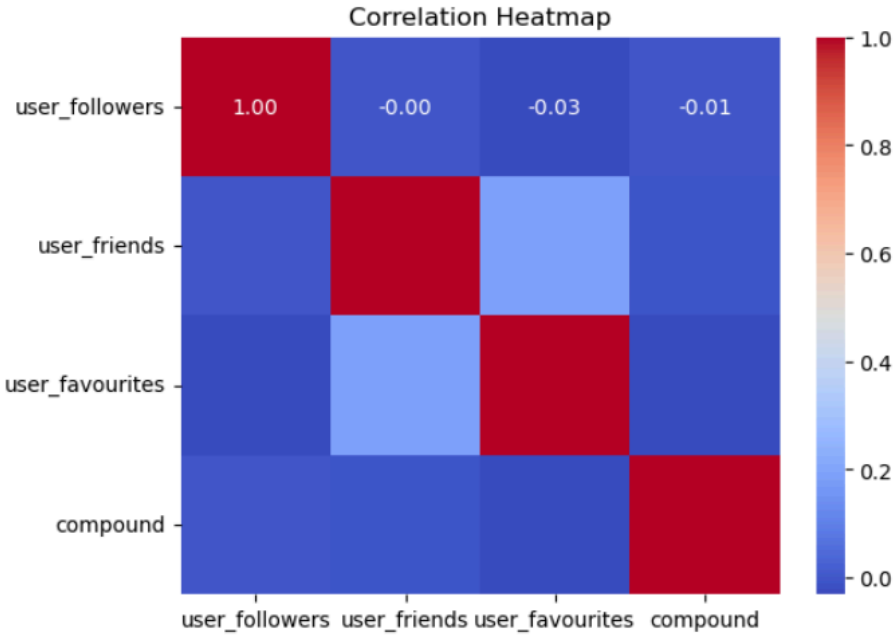


Figure 5f. Combined Emotion Radar Analysis

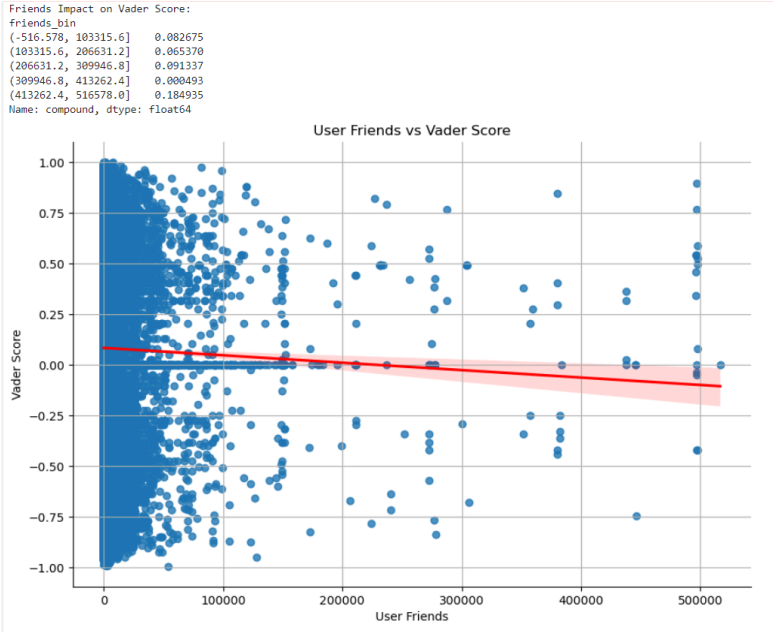


Figures 5g-h. Correlation Matrix and Heat Maps of User Attributes and VADER Scores

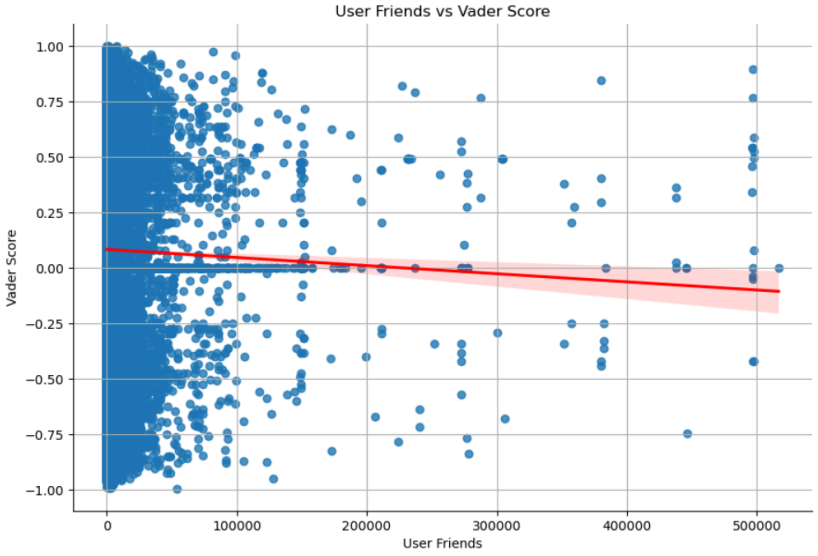
	user_followers	user_friends	user_favourites	compound
user_followers	1.000000	-0.002044	-0.028656	-0.005729
user_friends	-0.002044	1.000000	0.194532	-0.007123
user_favourites	-0.028656	0.194532	1.000000	-0.030846
compound	-0.005729	-0.007123	-0.030846	1.000000



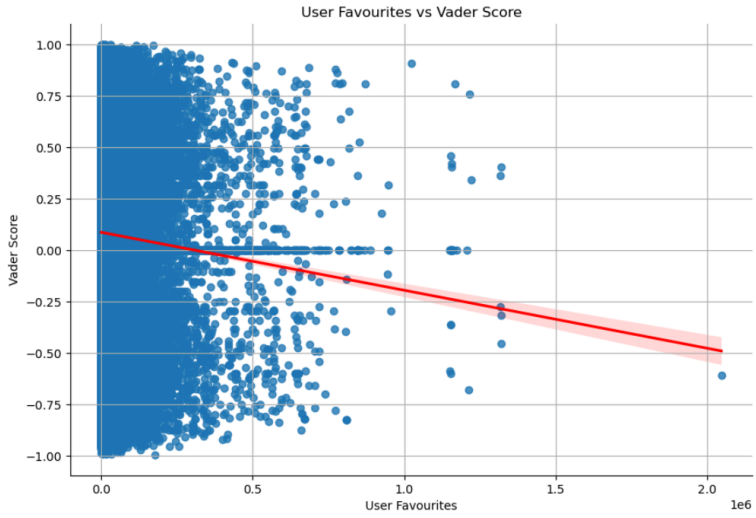
Figures 5i-5k. Linear Regression for User Attributes and VADER Sentiment Scores



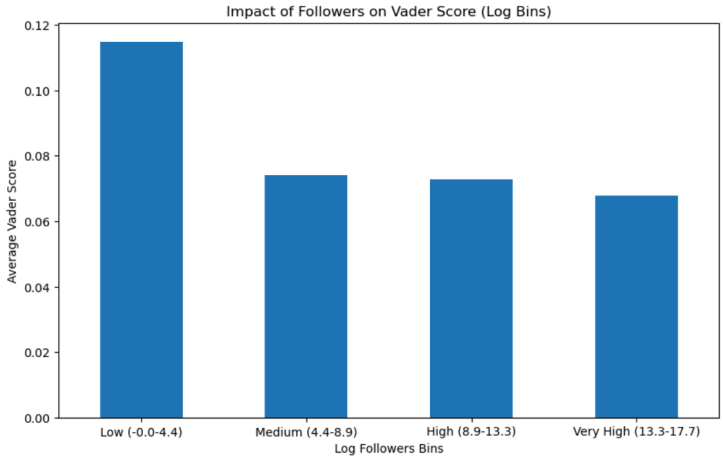
```
Friends Impact on Vader Score:
friends_bin
(-516.578, 103315.6]    0.082675
(103315.6, 206631.2]    0.065370
(206631.2, 309946.8]    0.091337
(309946.8, 413262.4]    0.000493
(413262.4, 516578.0]    0.184935
Name: compound, dtype: float64
```

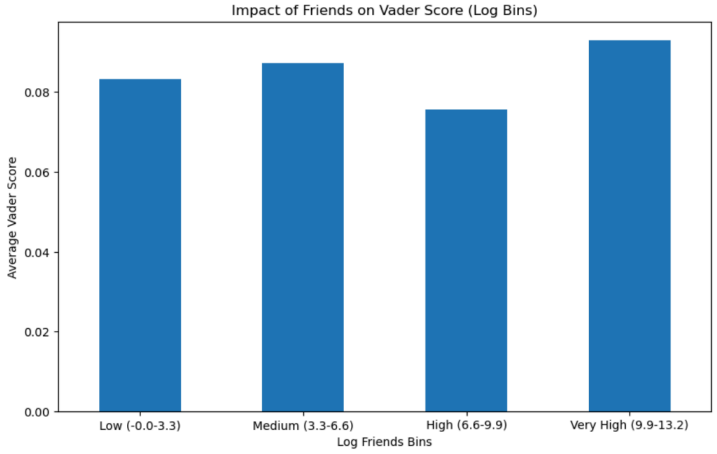



```
Favourites Impact on Vader Score:
favourites_bin
(-2047.197, 409439.4]     0.082864
(409439.4, 818878.8]     -0.020818
(818878.8, 1228318.2]     0.053839
(1228318.2, 1637757.6]   -0.056640
(1637757.6, 2047197.0]   -0.609300
Name: compound, dtype: float64
```



Figures 5l-n. Log Transformation and Quantile Analysis





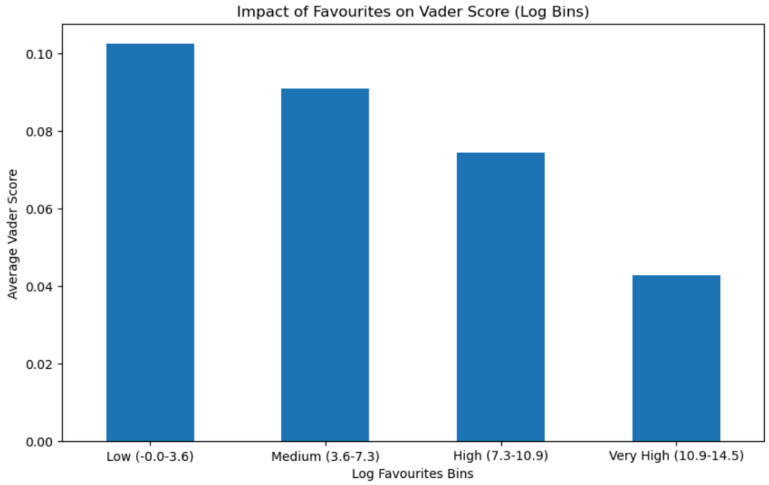


Figure 6a. PC1 Value of Top 20 Frequent Words

PC1 values of top 20 frequent words, where PC1 is the principal component obtained by applying PCA to the 768-dimensional word vectors derived from word embeddings; it represents the vector that best captures the sentiment direction in the reduced 2D space.

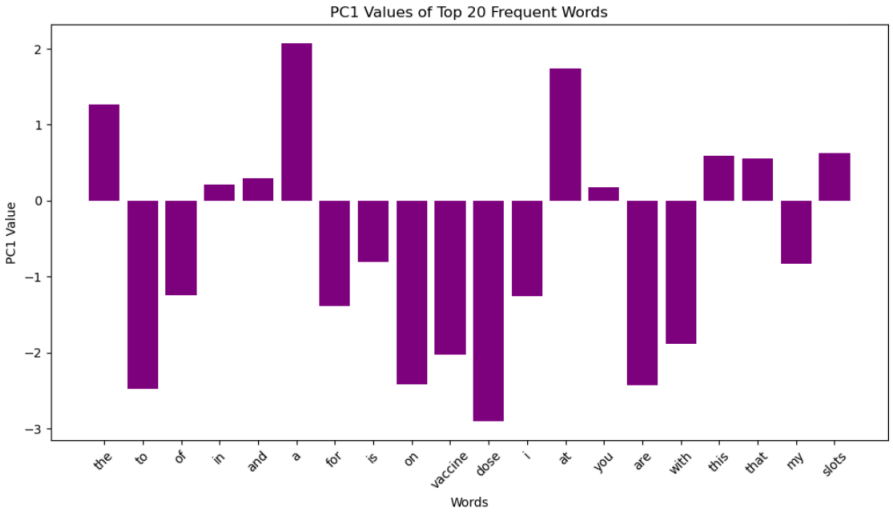


Figure 6b. PC1 Score vs VADER Compound Sentiment Scores