UNIVERSITY
of
INFORMATION TECHNOLOGY

CST-42315: Data Analysis and Management

Project Title

Vaccine Tweets Sentiment Analysis

Group Members

Aung Khant Myat [4KE-1000]
Tun Ye Minn [4KE-984]
Khin Nyo Nyo Theint [4KE - 970]
Youn Thinzar [4KE-1027]
Htet Aung Hlaing [4KE- 659]
Ye Thu Aung [4KE- 765]
Minn Khant Paing [4KE- 1169]
Khine Zin Thant [4KE- 1153]

Submitted Date: …9/14/2023…

Contents

Vaccine Tweets Sentiment Analysis

1.Introduction

The development and distribution of vaccines have become pivotal milestones in our collective journey toward normalcy in an era defined by the COVID-19 pandemic. Tweets become an important real-time hub for public discourse as COVID-19 vaccine discussions unfold across different platforms. With Natural Language Processing (NLP) and data analysis, our project explores this vast and dynamic landscape of Twitter for insights and sentiment patterns related to COVID-19 vaccinations. Sentiment analysis, often referred to as opinion mining, is a technique within natural language processing (NLP) that focuses on deciphering the emotional tone or sentiment expressed in textual data. It involves analyzing text and classifying it as either positive, negative, neutral, or somewhere along a spectrum of emotions.

1.1. Objective

With this project, we shed light on Twitter sentiment's rich tapestry and reveal valuable insights into public sentiment dynamics and COVID-19 vaccines' impact on digital discourse. Through analysis of sentiment expressed on Twitter, our project helps us understand how the general public views and feels about COVID-19 vaccines. Analyzing temporal sentiment trends over time can help identify when and how public sentiment has evolved in response to vaccine-related events, news, or policy changes. This can aid in trend prediction and preparation for future developments. Location-Specific Insights help pinpoint regions where sentiments are particularly strong, be it positive or negative. This information can guide targeted interventions or communication strategies in specific areas. User Behavior Analysis involve categorizing users by popularity and analyzing their sentiments can help in identifying key opinion leaders and understanding how different user groups engage with vaccine-related discussions. This knowledge can be useful for social media strategies. Policymakers can use sentiment analysis to gauge public reception of vaccination policies and make data-driven decisions to improve public health outcomes. Furthermore, our project can contribute to academic research on sentiment analysis, text mining, and the dynamics of public sentiment in response to health crises.

1.2. Project Flow

Starting with data retrieval and preprocessing, we efficiently handle large volumes of twitter data, cleaning and storing it for data analysis. Afterwards, we begin with a comprehensive exploration of Twitter sentiment analysis, seeking to unravel the diverse range of emotions and opinions expressed on the social media platform. Starting from an exploration of sentiment distribution within Twitter, we explore positivity, negativity, and neutrality as a spectrum of emotions. Afterward, we explore sentiment trends over time. A more detailed analysis is conducted to uncover the most emotional tweets, focusing on geographical areas that have a significant impact on tweet sentiment. As part of our analysis, we perform a Correlation Analysis, where we uncover the relationship between sentiment and external factors. A further investigation into user behavior is also conducted to identify how different categories of user's express sentiment about COVID-19 vaccines. As we navigate the intricacies of text data, Text Decomposition Analysis (TDA) is employed to reduce the dimensionality of the data and identify the key words that determine sentiment variance. As a final step, we perform a comprehensive sentiment analysis, calculating positive, negative, and neutral sentiments, along with polarity scores, and assessing its accuracy using machine learning.

2. About Dataset

Dataset is about the Covid19 Vaccines mainly about the Pfizer/Biotech Vaccine Tweets from Twitter API. This dataset is from Kaggle.com.

2.1. Data

The dataset includes 11020 rows and 17 columns. The data type includes a mixed of textual data, categorical data, and numerical data. The dataset is in json format.
The features of the dataset are:

- **id** : Twitter user's id
- **user_name** : Twitter user name
- **user_location** : Location of the twitter user posting the tweet
- **user_description** : Twitter user's Twitter Profile Information
- **user_created** : The date user create twitter account
- **user_follower** : Number of followers of the user
- **user_friends** : Number of friends the user have on Twitter
- **user_favourites** : Number of tweets the user favorited
- **user_verified** : Whether the user is verified profile or not

- **date** : The date where the tweet is posted
- **text** : Text on the tweet
- **hashtags** : Hashtag on the tweet
- **source** : Device the tweet is posted
- **Retweets** : Number of retweets
- **Favourites** : Number of time the tweet had been favorited
- **Is_retweet** : Whether the tweet is the retweet post of others

2.2. Nature of Text Data

In our sentiment analysis of Twitter data, we encountered a variety of sentiments:

1. **Negative**: "Five years ahead of schedule? It's a pandemic! There was no pre-determined schedule." This tweet expresses a negative sentiment, highlighting frustration and criticism regarding an event not going as planned during the pandemic.

2. **Neutral**: "Yes, here is the link to the patient education information from #PfizerBioNTec." This tweet conveys a neutral sentiment, merely providing a link to patient education information without expressing any strong emotions.

3. **Positive**: "Happy and relieved to have the Pfizer BioNTech Covid vaccine. #GetVaccinated!" This tweet radiates positivity and relief, with the user expressing happiness and encouraging others to get vaccinated, reflecting a positive sentiment.

2.3. How Data is retrieved?

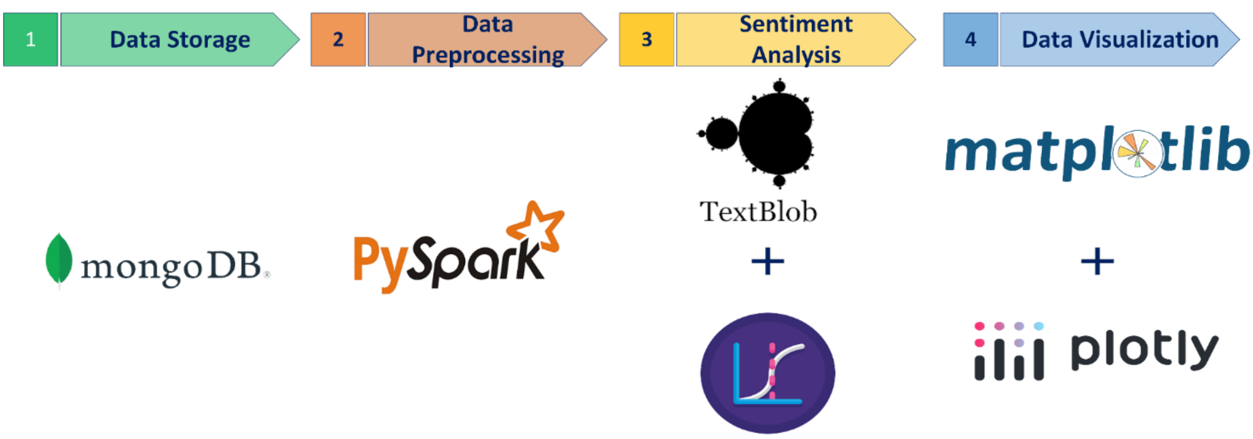
The dataset used for analysis is stored in **MongoDB Atlas**. Thus, the information is extracted from the database. MongoDB Atlas is used as the storage. By creating a free tier shared cluster (Cluster0), using AWS as cloud provider and choosing **N.Virginia(us-east-1)** region. The connection string is "**mongodb+srv://khinezinthant:(password)@cluster0.fkq2fra.mongodb.net/**". Thus, the dataset can be retrieved from teammates from different geographical locations to do separate analysis tasks.

3. Methodology

Sentiment analysis can be approached using various methods, including: rule-based approach, machine learning-based approach, and hybrid approach. For this analysis, we have chosen the machine learning-based approach.

3.1. System Architecture

The system architecture for the Vaccine Tweets Sentiment Analysis project includes data collection to gather relevant tweets, secure storage in MongoDB in JSON format, data preprocessing with PySpark for cleaning and transformation, and sentiment analysis using Text Blob and Logistic Regression. Matplotlib and Plotly are employed for data visualization, facilitating the creation of insightful visual representations. This architecture aims to provide valuable insights from tweet data while ensuring efficient data management and analysis.



System Components of the Architecture are:

- Data Storage
  - Purpose: Store tweet data in JSON format.
  - Technologies Used: MongoDB Atlas.
- Data Preprocessing
  - Purpose: Clean and transform raw data.
  - Technologies Used: PySpark.
- Sentiment Analysis
  - Purpose: Perform sentiment analysis on tweets.
  - Technologies Used: Text Blob, Logistic Regression.
- Data Visualization
  - Purpose: Create visual representations of data.
  - Technologies Used: Matplotlib, Plotly.

4. Sentiment Analysis Approach

Step-by-step Approach

Our sentiment analysis approach involves the following steps:

1. Data Preprocessing
2. Accessing sentiment using polarity function of TextBlob
3. Feature extraction
4. Splitting Training and Testing Dataset
5. Applying Machine Learning Models
6. Fine-tuning Parameters
7. Evaluation Metrics

Let's delve deeper into the step-by-step process used in our sentiment analysis.

4.1. Data Preprocessing

By using pyspark as a data cleaning tool, we clean the data to be used for our project. We convert the texts to lower case to better process data by using lower () function. We clear out the special characters that add no value to text-understanding and induce noise into algorithms by using regexp_replace ().

1. **Lower-casing**: Lower-casing converts all text to lowercase to ensure uniformity in text analysis.

2. **Removing noise from tweet text using Regex**: Noise removal involves eliminating irrelevant characters, symbols, and special characters (example- @,#,hyperlinks) from text using regular expressions.

3. **Tokenization**: Tokenization is a fundamental step in text data preprocessing that involves breaking down text into individual units, typically words or phrases, referred to as tokens. Tokens are the building blocks of text analysis, allowing for the isolation and examination of individual elements within a given text.

4. **Stop word removal**: In NLP, stop words are common words like "the," "and," "in," or "is" that are removed from text during NLP analysis because they don't contribute much to understanding the text's meaning. They are considered noise words that can be safely ignored to focus on the more important words in a document or sentence. By removing stop words, NLP algorithms can improve the efficiency and accuracy of tasks like text classification, sentiment analysis, and information retrieval. In our project, we used stop words from NLTK library. NLTK (Natural Language Toolkit) is

a popular Python library for NLP that provides a predefined list of stop words for various languages.

5. **Duplicate removal**: Duplicate removal is a critical step in data preprocessing that focuses on enhancing data quality. In the context of sentiment analysis, duplicate text entries can skew the analysis results and lead to biased models.

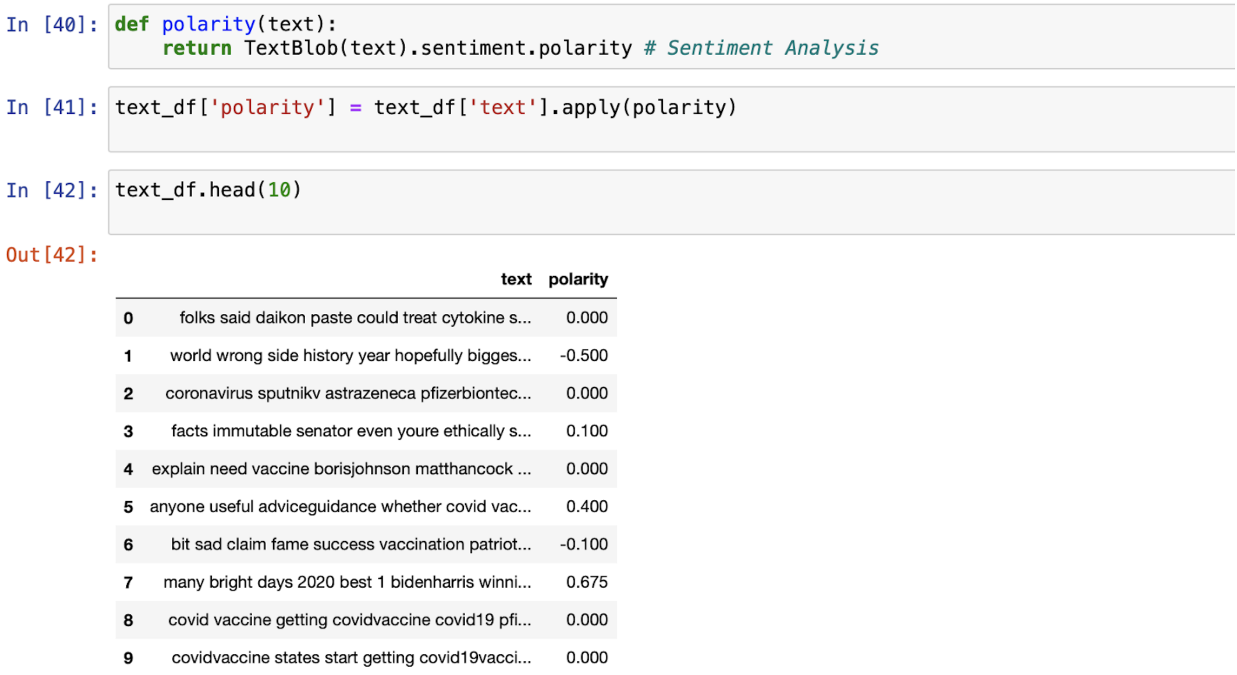
6. **Stemming**: The PorterStemmer is used for stemming. It is a popular stemming algorithm in natural language processing (NLP) and is available in Python through libraries like NLTK (Natural Language Toolkit). Stemming is a text normalization process that reduces words to their base or root form, often by removing suffixes. The purpose is to group words with the same root or meaning together, even if they are different grammatical forms of the word. In this sentiment analysis, we used it for TruncatedSVD in order to process the text data with ease and not for processing functionality.

7. **Lemmatization**: We used WordNetLemmatizer to perform lemmatization, which is the process of reducing words to their base or dictionary form, known as lemmas. Lemmatization is a more precise form of text normalization compared to stemming. Like PortStemmer, we used it for TruncatedSVD in order to process the text data with ease and not for processing functionality.

4.2. Accessing Sentiment Using Polarity Function of TextBlob

TextBlob is a Python library for natural language processing that includes a sentiment analysis feature. Sentiment polarity in TextBlob refers to the sentiment or emotional tone conveyed in a piece of text, and it is usually classified as either positive, negative, or neutral.

Polarity Score: Polarity is a critical aspect of sentiment analysis. The polarity score is a single numerical value that indicates the sentiment polarity of the text. It is measured using the 'sentiment. polarity' attribute from TextBlob, which provides a polarity score ranging from -1 (very negative) to 1 (very positive), with 0 indicating a neutral sentiment.



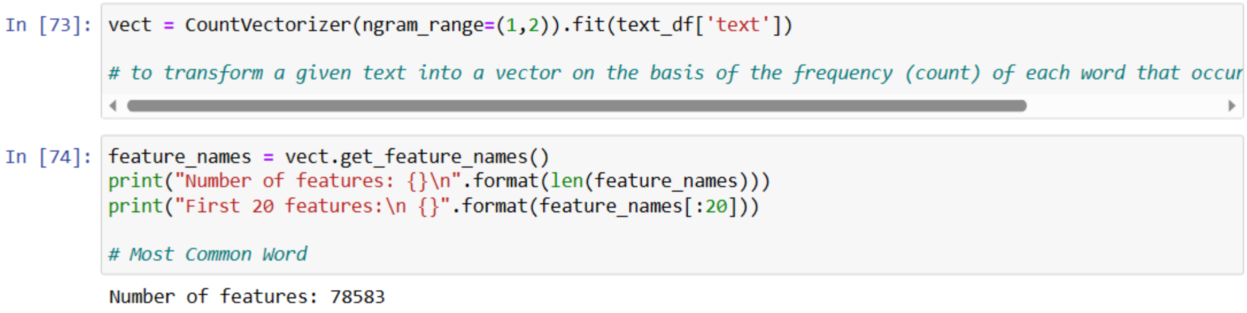
```
In [40]: def polarity(text):
             return TextBlob(text).sentiment.polarity # Sentiment Analysis

In [41]: text_df['polarity'] = text_df['text'].apply(polarity)

In [42]: text_df.head(10)

Out[42]:
```

	text	polarity
0	folks said daikon paste could treat cytokine s...	0.000
1	world wrong side history year hopefully bigges...	-0.500
2	coronavirus sputnikv astrazeneca pfizerbiontec...	0.000
3	facts immutable senator even youre ethically s...	0.100
4	explain need vaccine borisjohnson matthancock ...	0.000
5	anyone useful adviceguidance whether covid vac...	0.400
6	bit sad claim fame success vaccination patriot...	-0.100
7	many bright days 2020 best 1 bidenharris winni...	0.675
8	covid vaccine getting covidvaccine covid19 pfi...	0.000
9	covidvaccine states start getting covid19vacci...	0.000

4.3. Feature Extraction

For vectorization, Count Vectorizer is used to convert text data into numerical features that machine learning models can understand. It works by transforming a collection of text documents into a matrix of token counts, where each row represents a document, and each column represents a unique word or token in the entire corpus of documents. The values in the matrix indicate how many times each word occurs in each document.

In our implementation, ngram_range (1,2) is assigned in order to consider both unigrams (individual words) and bigrams (pairs of adjacent words) when extracting features from the text data.
As a result, the total of 78583 features is extracted from our data.



```
In [73]: vect = CountVectorizer(ngram_range=(1,2)).fit(text_df['text'])

         # to transform a given text into a vector on the basis of the frequency (count) of each word that occur

In [74]: feature_names = vect.get_feature_names()
         print("Number of features: {}\n".format(len(feature_names)))
         print("First 20 features:\n {}".format(feature_names[:20]))

         # Most Common Word

         Number of features: 78583
```

4.4. Splitting Training and Testing Dataset

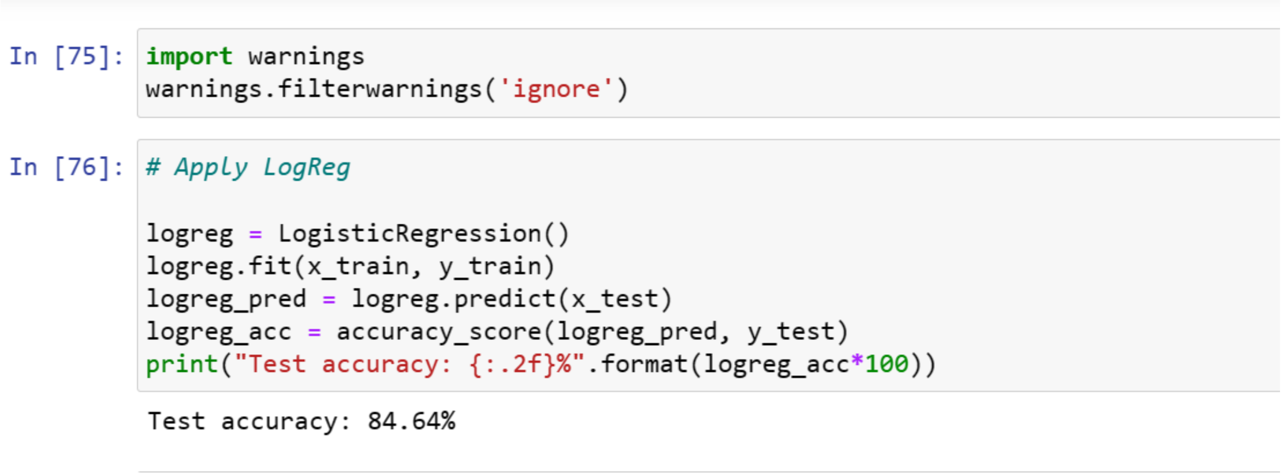
The dataset is divided into **training and testing subsets** (typically **80%-20%**) to evaluate model performance effectively.

4.5. Applying Machine Learning Models

Since our dataset is cleaned and preprocessed, we employ the following supervised learning algorithms as mentioned.

4.5.1. Logistic Regression

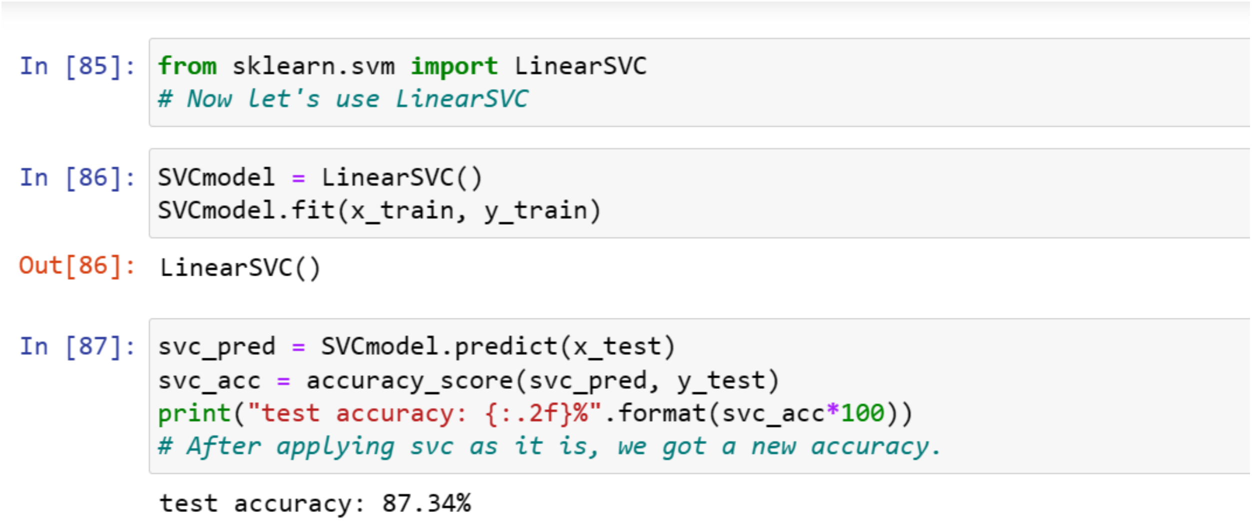
Logistic Regression is a supervised machine learning algorithm. Despite its name, logistic regression is a classification algorithm, not a regression algorithm. It models the relationship between a dependent variable and one or more independent features by estimating the probability that an input belongs to a particular class.



```
In [75]: import warnings
         warnings.filterwarnings('ignore')

In [76]: # Apply LogReg

         logreg = LogisticRegression()
         logreg.fit(x_train, y_train)
         logreg_pred = logreg.predict(x_test)
         logreg_acc = accuracy_score(logreg_pred, y_test)
         print("Test accuracy: {:.2f}%".format(logreg_acc*100))

         Test accuracy: 84.64%
```

4.5.2. Support Vector Machine (SVM)

Another machine learning algorithm, Linear Support Vector Classifier (LinearSVC) is used for sentiment analysis. LinearSVC is a type of Support Vector Machine (SVM) that works well for text classification tasks, including sentiment analysis. It tries to find a hyperplane that best separates the data into different sentiment classes. The LinearSVC model is trained on the training data. During training, the model learns the relationships between the extracted features and the corresponding sentiment labels. After training, the model is tested on the testing set to assess its accuracy and performance by confusion matrix.



```
In [85]: from sklearn.svm import LinearSVC
         # Now let's use LinearSVC

In [86]: SVCmodel = LinearSVC()
         SVCmodel.fit(x_train, y_train)

Out[86]: LinearSVC()

In [87]: svc_pred = SVCmodel.predict(x_test)
         svc_acc = accuracy_score(svc_pred, y_test)
         print("test accuracy: {:.2f}%".format(svc_acc*100))
         # After applying svc as it is, we got a new accuracy.

         test accuracy: 87.34%
```

4.6. Fine-tuning Parameters

Grid search is a fundamental technique for hyperparameter tuning in machine learning models like Support Vector Classifier (SVC) and Logistic Regression (LogReg) used in sentiment analysis. Hyperparameters, such as C (regularization strength), kernel type, degree (for polynomial kernels), and gamma (kernel coefficient), profoundly influence model performance. Grid search systematically explores predefined sets of hyperparameter values, like [0.01, 0.1, 1, 10] for C and ["linear", "poly", "rbf", "sigmoid"] for the kernel, training and evaluating models through cross-validation. The aim is to pinpoint the hyperparameter combination that delivers the best performance, often assessed by metrics like accuracy or F1-score. Once identified, these optimal hyperparameters are applied to train final models, ensuring they are finely tuned for the sentiment analysis task. Grid search simplifies the intricate process of optimizing hyperparameters, serving as an indispensable tool for enhancing model accuracy.

Let's explain how grid search works for both the Logistic Regression and Support Vector Machine (SVM) models used in sentiment analysis:

4.6.1. Grid Search for Logistic Regression

1. **Select Hyperparameters**: Identify the hyperparameters that need tuning for the logistic regression model. In the case of logistic regression, common hyperparameters include the regularization parameter (C), penalty (L1 or L2), and solver.

2. **Define Hyperparameter Grid**: Specify a grid of possible values for each hyperparameter. For example:
    a. C: [0.01, 0.1, 1, 10]
    b. Penalty: ['l1', 'l2']
    c. Solver: ['liblinear', 'lbfgs']

3. **Grid Search Process**: Grid search will create combinations of hyperparameters (e.g., C=0.01, Penalty='l1', Solver='liblinear') and evaluate the model's performance using a specified evaluation metric (e.g., accuracy, F1-score) through cross-validation.

4. **Cross-Validation**: For each combination of hyperparameters, cross-validation is performed. The dataset is split into training and validation subsets multiple times (usually k-fold cross-validation), and the model's performance is averaged over these folds to reduce overfitting.

5. **Evaluation Metric**: The chosen evaluation metric (e.g., accuracy) is used to assess the model's performance for each combination of hyperparameters.

6. **Select Best Hyperparameters**: After evaluating all combinations, grid search identifies the set of hyperparameters that resulted in the best performance based on the chosen evaluation metric.

7. **Model Training**: Finally, the model is trained using the entire training dataset with the best hyperparameters identified by grid search.

4.6.2. Grid Search for Support Vector Machine (SVM)

1. **Select Hyperparameters**: Identify the hyperparameters that need tuning for the SVM model. For SVM, common hyperparameters include the kernel (linear, polynomial, radial basis function, etc.), C (regularization parameter), and gamma (kernel coefficient).

2. **Define Hyperparameter Grid**: Specify a grid of possible values for each hyperparameter. For example:
    - Kernel: ['linear', 'poly', 'rbf']
    - C: [0.1, 1, 10]
    - Gamma: [0.01, 0.1, 1]

3. **Grid Search Process**: Grid search will create combinations of hyperparameters (e.g., Kernel='linear', C=0.1, Gamma=0.01) and evaluate the SVM model's performance using cross-validation.

4. **Cross-Validation**: Similar to logistic regression, cross-validation is performed for each combination of hyperparameters to ensure robust evaluation.

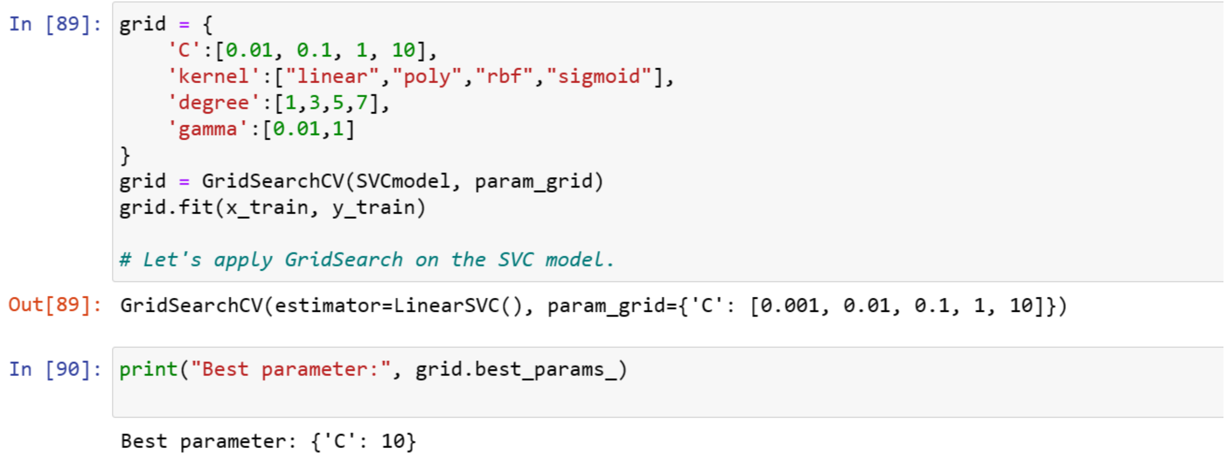
5. **Evaluation Metric**: The chosen evaluation metric (e.g., accuracy, F1-score) is used to assess the SVM model's performance for each combination of hyperparameters.

6. **Select Best Hyperparameters**: Grid search identifies the set of hyperparameters that

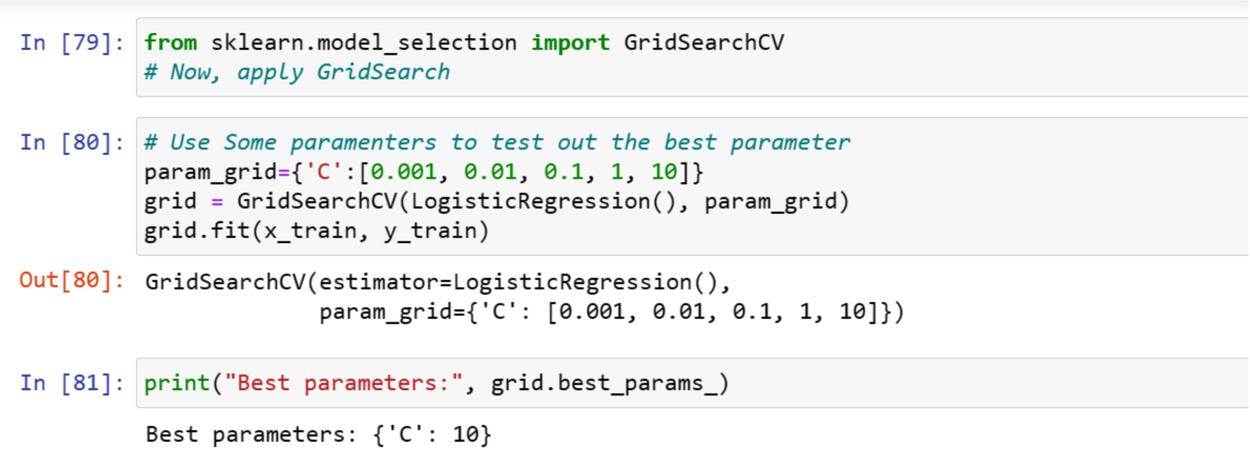
yielded the best performance based on the chosen evaluation metric.

7. **Model Training**: The SVM model is trained using the entire training dataset with the best hyperparameters.

Grid search automates the process of hyperparameter tuning and helps find the hyperparameters that optimize the model's performance, ensuring that the model generalizes well to new data. It prevents the need for manual tuning, which can be time-consuming and less systematic.



```
In [89]: grid = {
             'C':[0.01, 0.1, 1, 10],
             'kernel':["linear","poly","rbf","sigmoid"],
             'degree':[1,3,5,7],
             'gamma':[0.01,1]
         }
         grid = GridSearchCV(SVCmodel, param_grid)
         grid.fit(x_train, y_train)

         # Let's apply GridSearch on the SVC model.

Out[89]: GridSearchCV(estimator=LinearSVC(), param_grid={'C': [0.001, 0.01, 0.1, 1, 10]})

In [90]: print("Best parameter:", grid.best_params_)

         Best parameter: {'C': 10}
```



```
In [79]: from sklearn.model_selection import GridSearchCV
         # Now, apply GridSearch

In [80]: # Use Some paramenters to test out the best parameter
         param_grid={'C':[0.001, 0.01, 0.1, 1, 10]}
         grid = GridSearchCV(LogisticRegression(), param_grid)
         grid.fit(x_train, y_train)

Out[80]: GridSearchCV(estimator=LogisticRegression(),
                      param_grid={'C': [0.001, 0.01, 0.1, 1, 10]})

In [81]: print("Best parameters:", grid.best_params_)

         Best parameters: {'C': 10}
```

4.7. Evaluation Metrics

Evaluation metrics are crucial for assessing the performance of sentiment analysis models. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These values are used to calculate various performance metrics, including precision, recall, and the F1-score as follows:

Accuracy: (TP + TN) / (TP + TN + FP + FN)
Precision: TP / (TP + FP)
Recall (Sensitivity or True Positive Rate): TP / (TP + FN)
Specificity (True Negative Rate): TN / (TN + FP)
F1-Score: 2 * (Precision * Recall) / (Precision + Recall)

4.7.1. Accuracy

Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or there is no class imbalance.

4.7.2. Precision

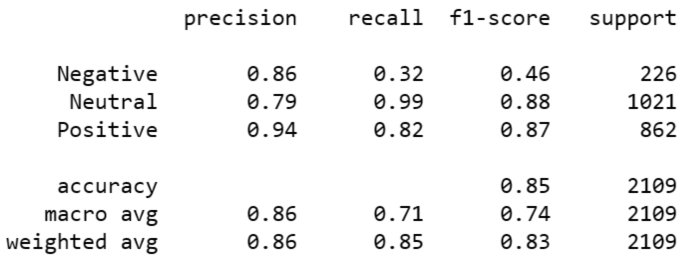
Precision is defined as the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes. Precision is a useful metric in cases where False Positive is a higher concern than False Negatives.

4.7.3. Recall

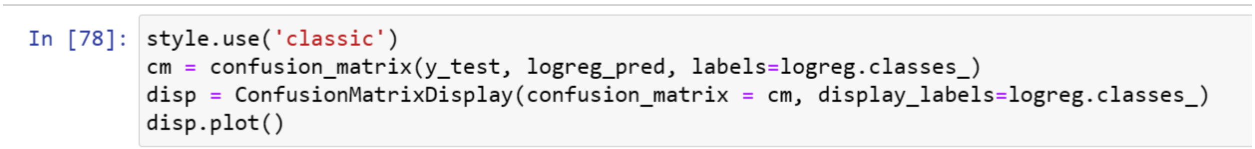
Recall is defined as the ratio of the total number of correctly classified positive classes divided by the total number of positive classes. Or, out of all the positive classes, how much we have predicted correctly. Recall should be high (ideally 1). Recall is a useful metric in cases where False Negative trumps False Positive.

4.7.4. F1-Score

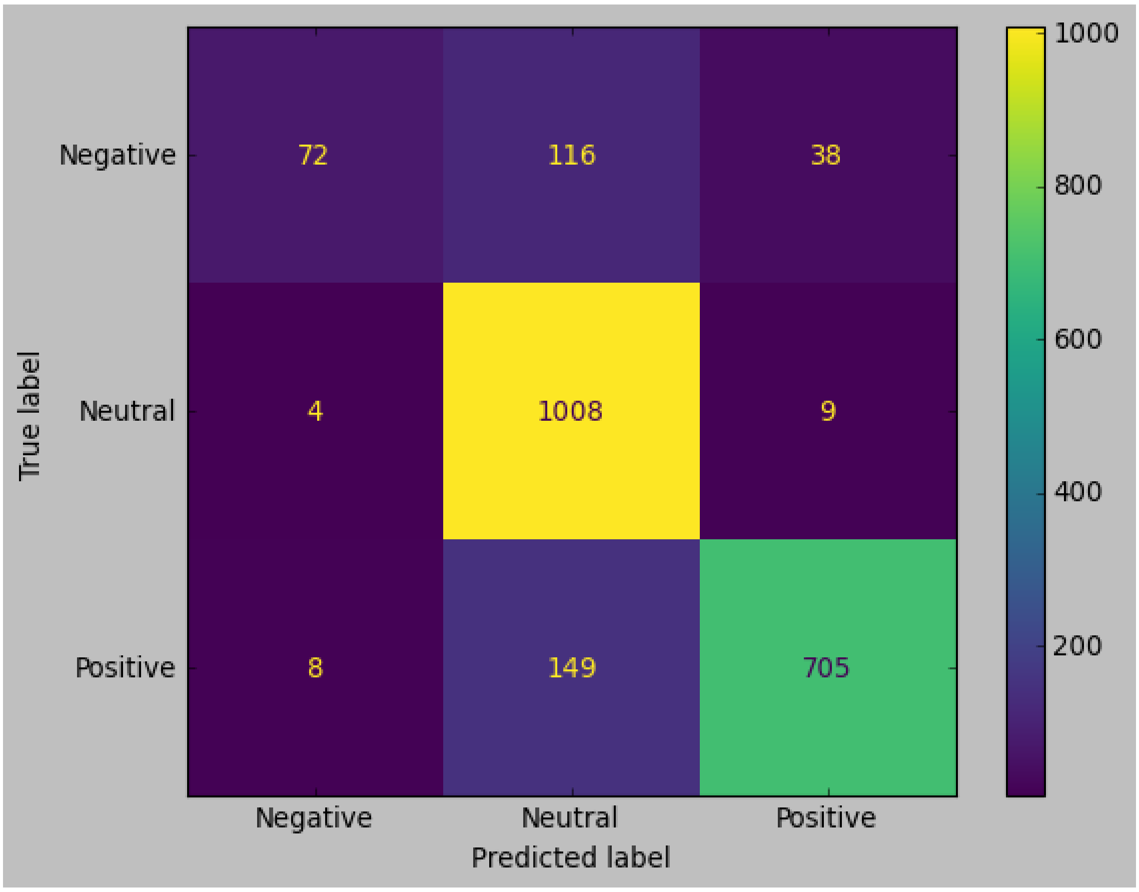
The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall.



```
              precision    recall  f1-score   support

    Negative       0.86      0.32      0.46       226
     Neutral       0.79      0.99      0.88      1021
    Positive       0.94      0.82      0.87       862

    accuracy                           0.85      2109
   macro avg       0.86      0.71      0.74      2109
weighted avg       0.86      0.85      0.83      2109
```

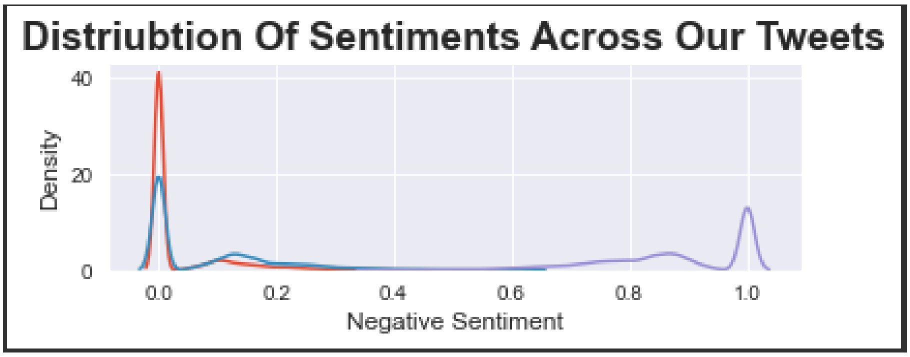


```
In [78]: style.use('classic')
         cm = confusion_matrix(y_test, logreg_pred, labels=logreg.classes_)
         disp = ConfusionMatrixDisplay(confusion_matrix = cm, display_labels=logreg.classes_)
         disp.plot()
```



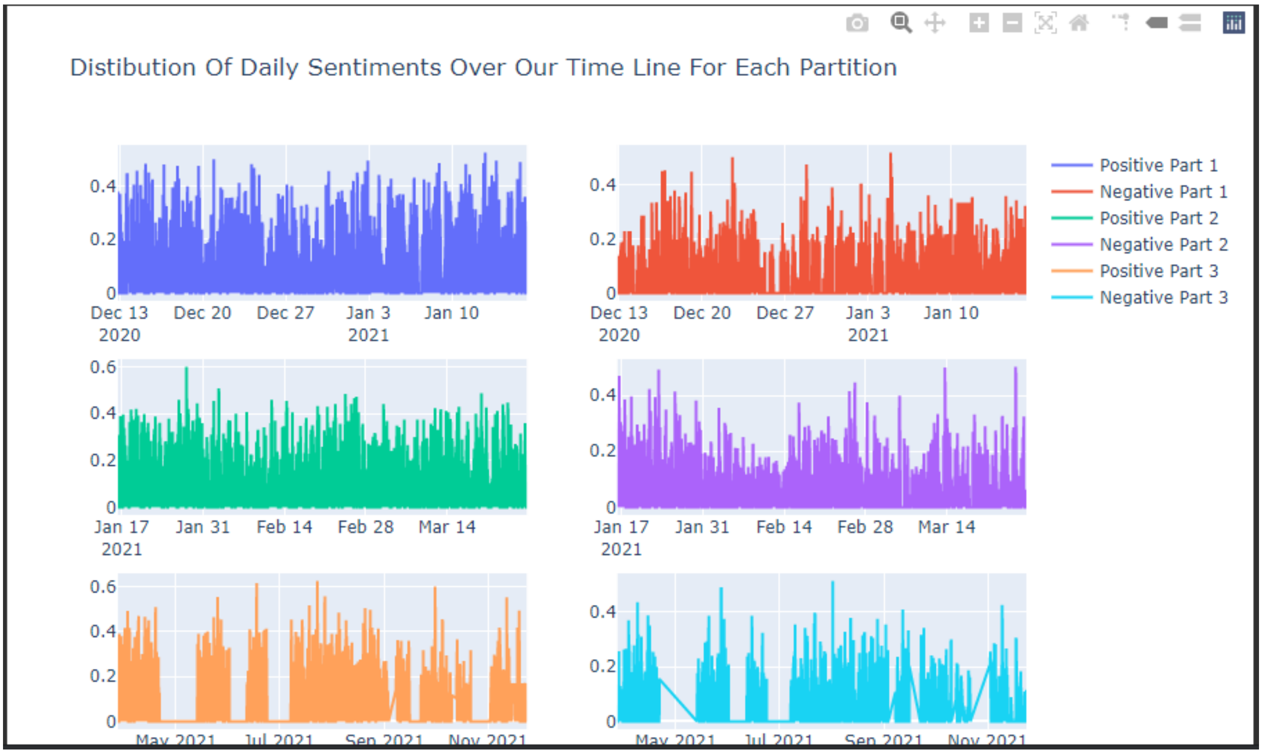
5. Insight

5.1. Distribution of Sentiment Data



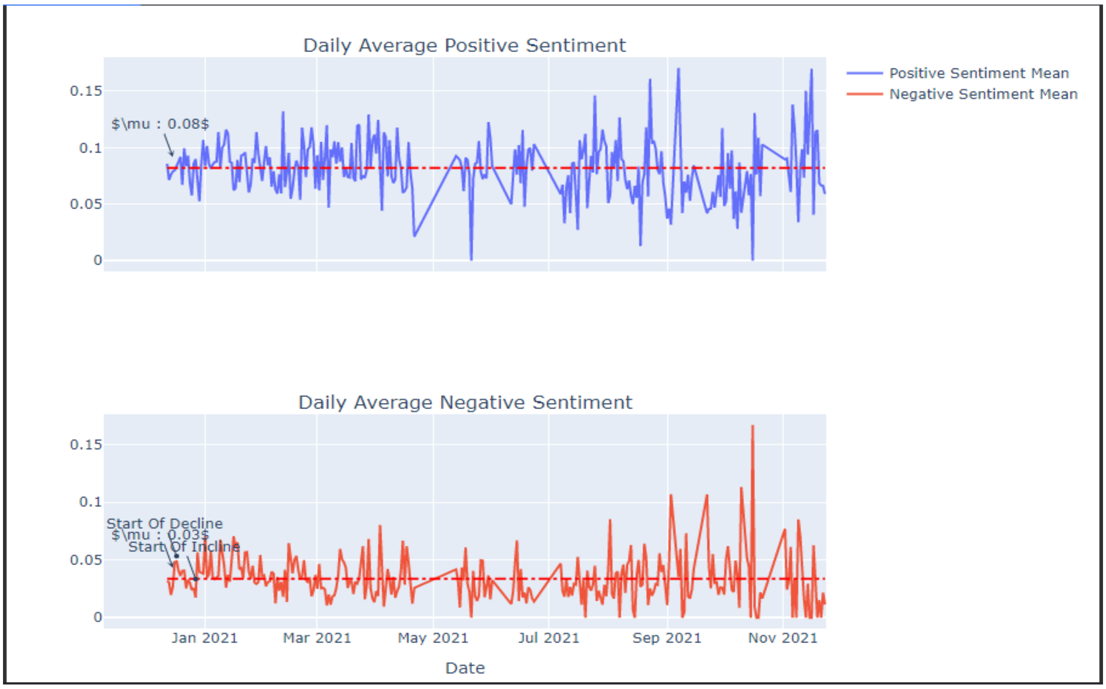
You can observe that the distributions of the sentiments follow a normal distribution; the negative and positive sentiments are very similar, proposing that there may be no significant differences in the strength of our data's positive and negative sentiments. It is also clear that the dominant sentiment is neutral; oddly, most of the tweets do not resemble more positive or negative sentiment rather neutral.

5.2. Time Based Analysis



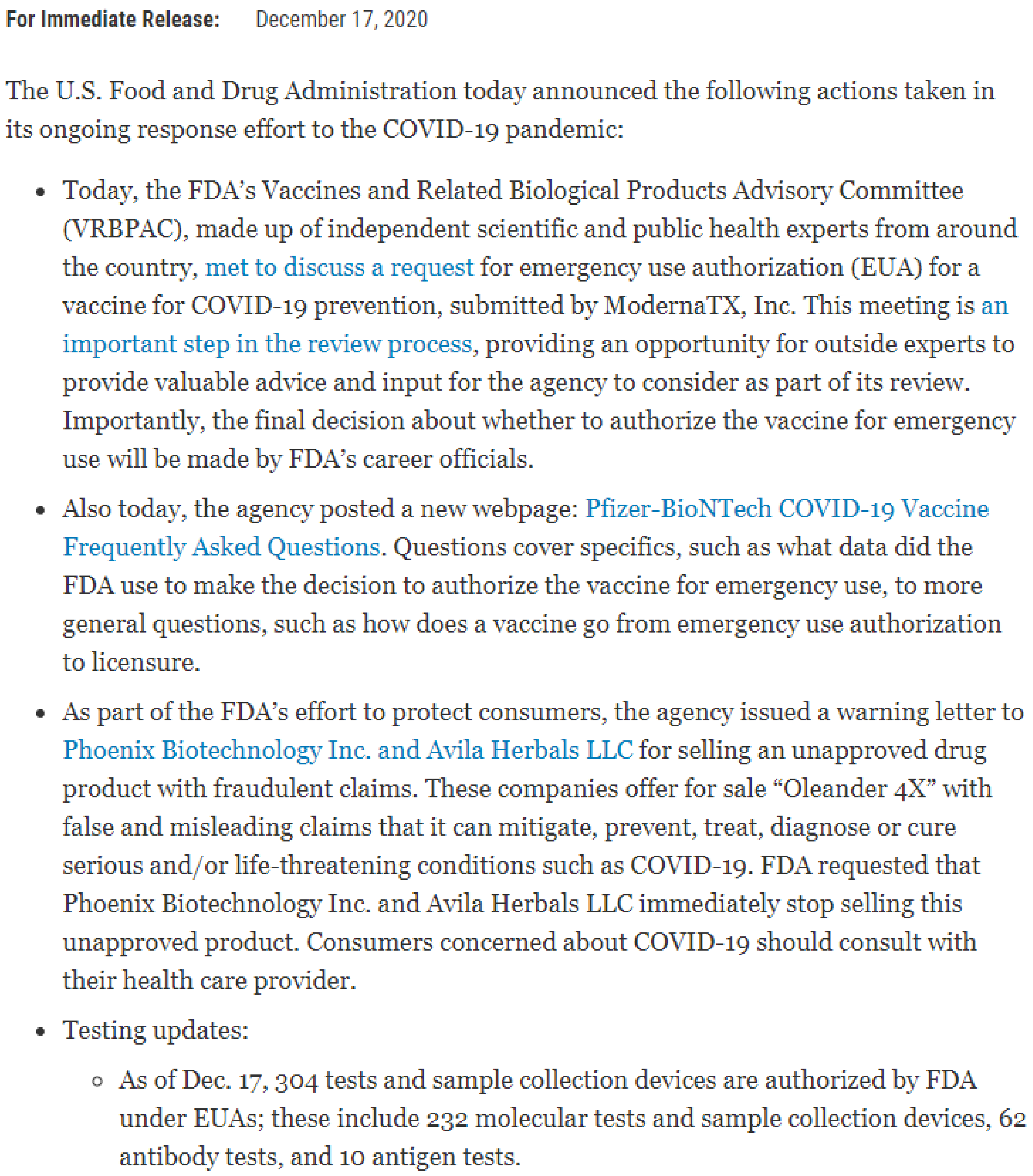
We see that the tweets' sentiments do not meet stationarity requirements as to non-constant mean and variance. In the above code cell, we have tested our hypothesis on 3 partitions of our data. Might it indicate we have some trend in our data?

5.3. Sentiment Average Change with time



One observation that interests me the most is that from the 20 of December to the 27 of December, there is a decline in the strength of the average negative sentiment; what happened in that week that induced this decline?

What happened on the 17th December?

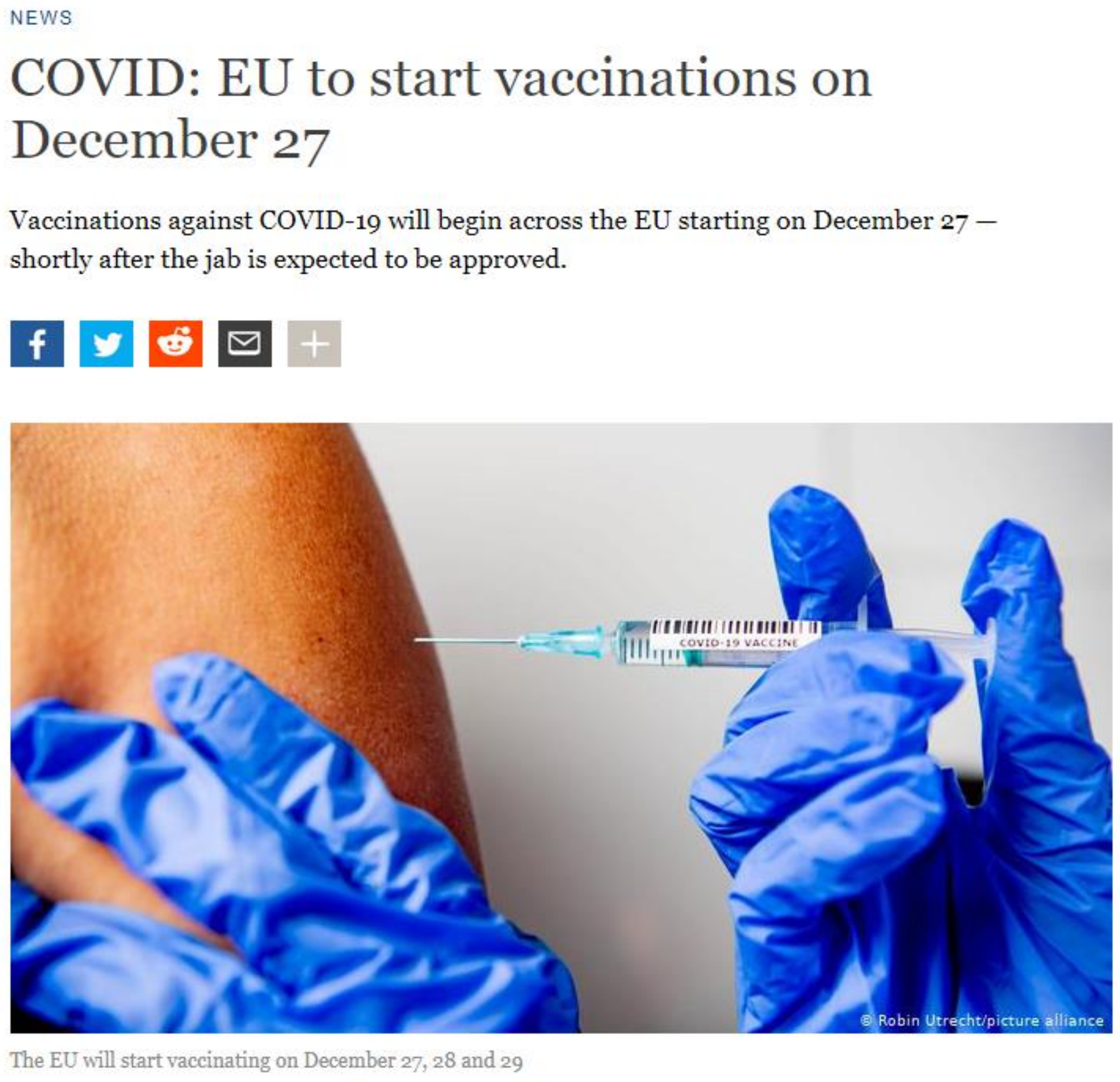


For Immediate Release: December 17, 2020

The U.S. Food and Drug Administration today announced the following actions taken in its ongoing response effort to the COVID-19 pandemic:

- Today, the FDA's Vaccines and Related Biological Products Advisory Committee (VRBPAC), made up of independent scientific and public health experts from around the country, met to discuss a request for emergency use authorization (EUA) for a vaccine for COVID-19 prevention, submitted by ModernaTX, Inc. This meeting is an important step in the review process, providing an opportunity for outside experts to provide valuable advice and input for the agency to consider as part of its review. Importantly, the final decision about whether to authorize the vaccine for emergency use will be made by FDA's career officials.
- Also today, the agency posted a new webpage: Pfizer-BioNTech COVID-19 Vaccine Frequently Asked Questions. Questions cover specifics, such as what data did the FDA use to make the decision to authorize the vaccine for emergency use, to more general questions, such as how does a vaccine go from emergency use authorization to licensure.
- As part of the FDA's effort to protect consumers, the agency issued a warning letter to Phoenix Biotechnology Inc. and Avila Herbals LLC for selling an unapproved drug product with fraudulent claims. These companies offer for sale "Oleander 4X" with false and misleading claims that it can mitigate, prevent, treat, diagnose or cure serious and/or life-threatening conditions such as COVID-19. FDA requested that Phoenix Biotechnology Inc. and Avila Herbals LLC immediately stop selling this unapproved product. Consumers concerned about COVID-19 should consult with their health care provider.
- Testing updates:
  - As of Dec. 17, 304 tests and sample collection devices are authorized by FDA under EUAs; these include 232 molecular tests and sample collection devices, 62 antibody tests, and 10 antigen tests.

Possible Explanation for The Incline on the 27th



NEWS

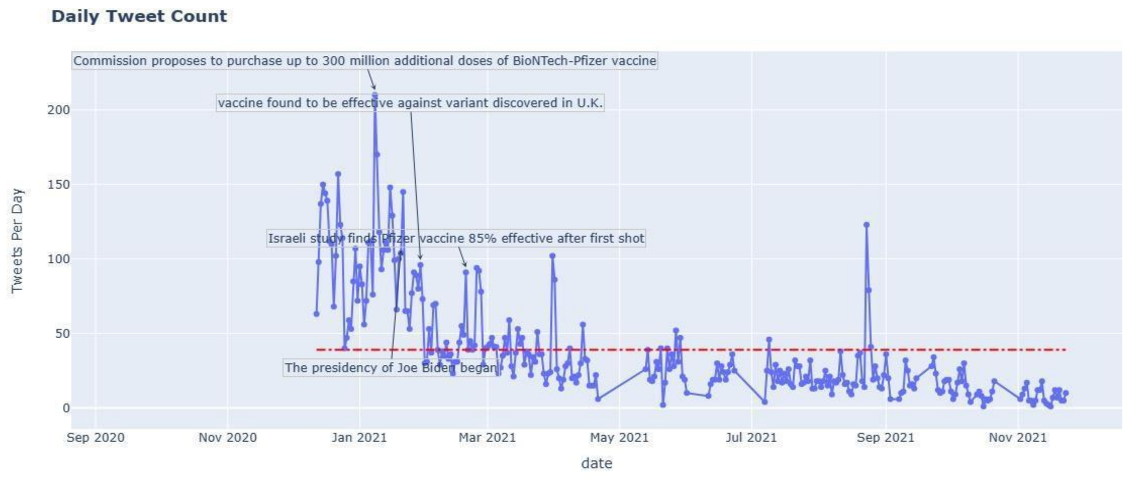
COVID: EU to start vaccinations on December 27

Vaccinations against COVID-19 will begin across the EU starting on December 27 — shortly after the jab is expected to be approved.

© Robin Utrecht/picture alliance

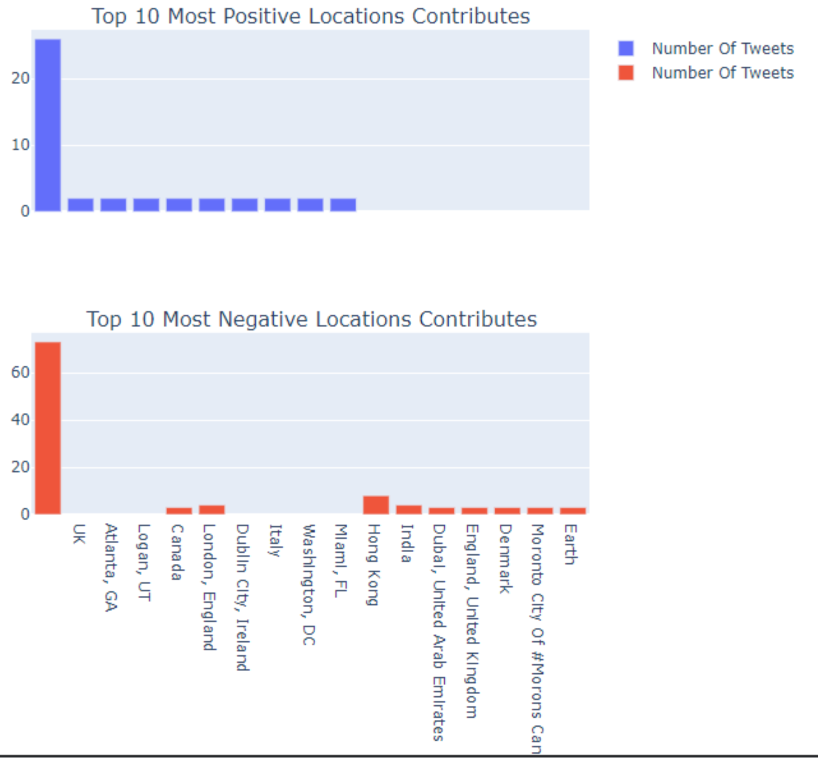
The EU will start vaccinating on December 27, 28 and 29

5.4. Daily Tweet Count



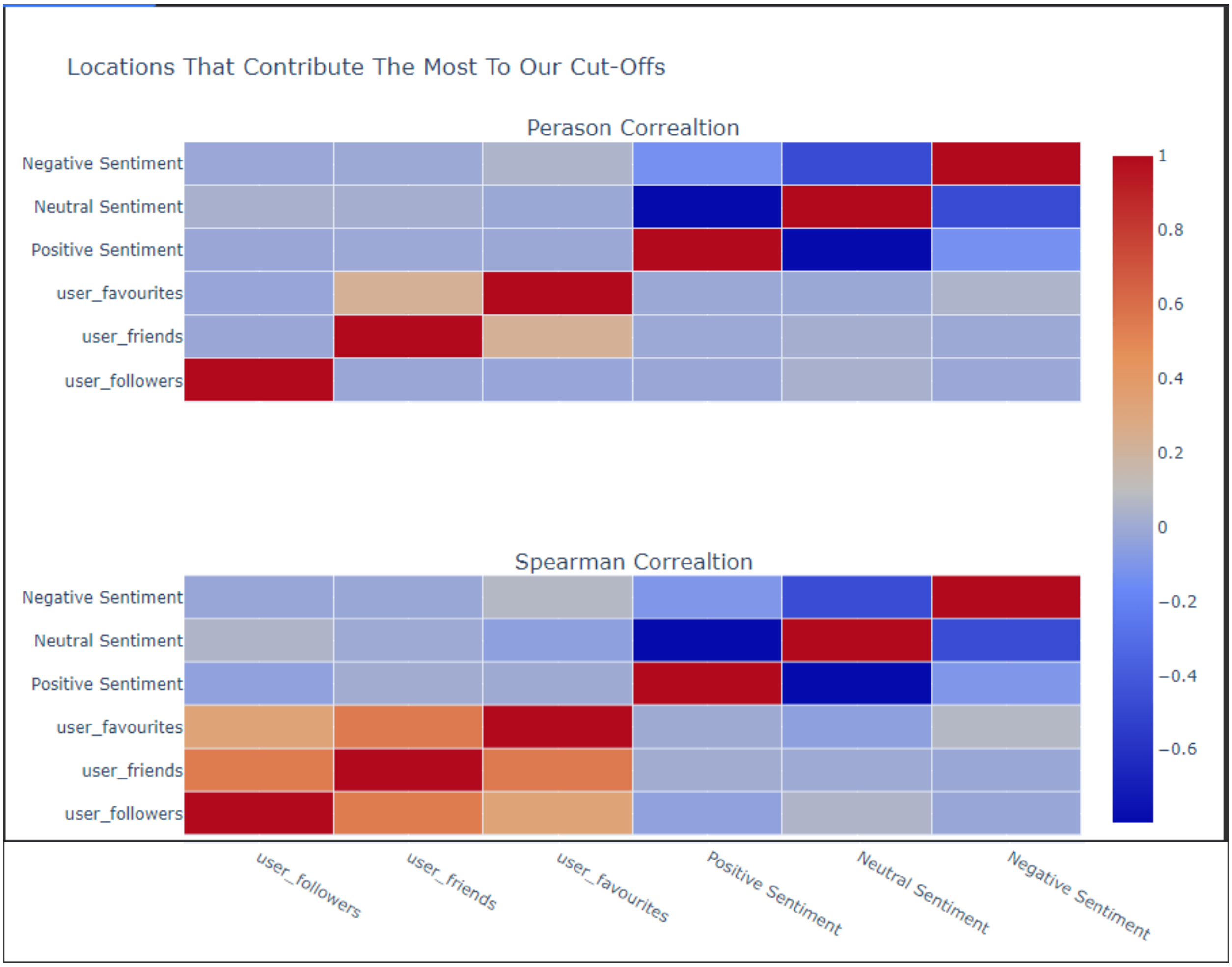
5.5. Extreme Sentiment Analysis

Locations That Contribute the Most to Our Cut-offs.



London and Miami are the leading locations when looking at the tweets we took from our positive cut-off in contrast to Moronto City, which stands out in the most negative cut-off.

5.6. Correlation Analysis



Unfortunately, we see no significant correlation between the tweet sentiments and any other numeric features given in our dataset, especially those that describe users.

5.7. User Analysis

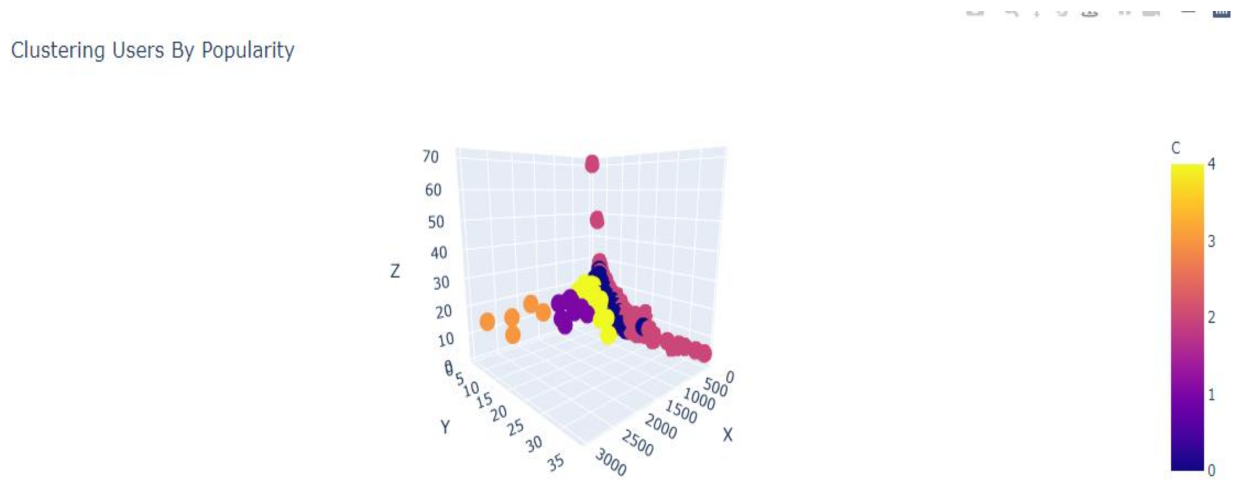
This section of our analysis will define a new metric representing a user's popularity; we will define a formula that will output a score of popularity based on the user's follower, friend, and favorite counts. The new metric will allow us to analyze the different users and the contribution to the average sentiments by different groups of user popularity. We will define user popularity using the following formula:

$$\Theta = \left\{ \begin{matrix} 1 \text{ if user is verified} \\ \text{else } 0.5 \end{matrix} \right\}$$

$$\text{Popularity:}\rho$$

$$= \sqrt{0.65 * \text{User Followers} + 0.25 * \text{User Friends} + 0.10 * \text{User Favourites}}$$

$$* \Theta$$

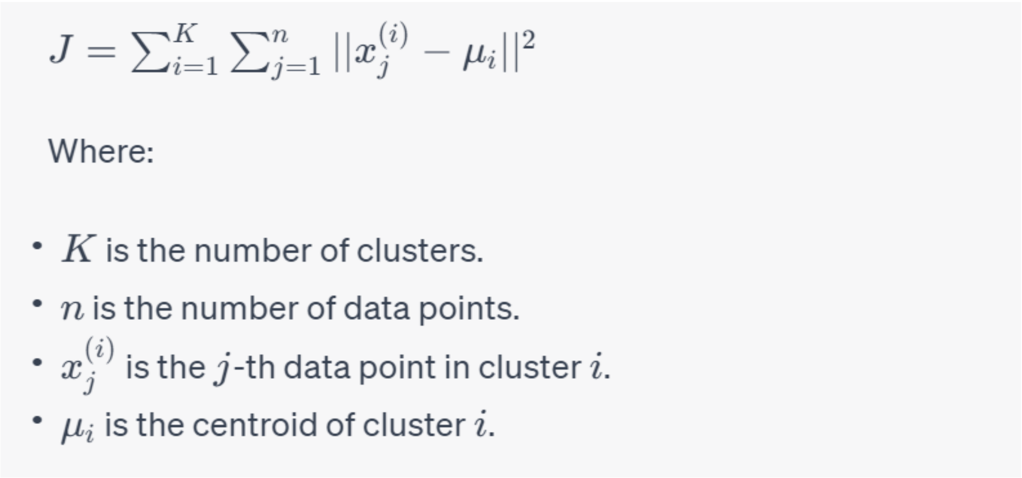


5.7.1. K-Means

The primary purpose of K-Means clustering is to categorize users into distinct groups based on their popularity scores. This segmentation allows for the exploration of user behavior patterns and facilitates data-driven decision-making by enabling the analysis of user engagement and preferences within each cluster. The resulting clusters provide valuable insights into how users with different levels of popularity interact with the content.

Why is K-mean used in user analysis of twitter data?

Twitter has a massive and diverse user base, and K-means clustering can help segment users into distinct groups based on various factors like tweet content, engagement patterns, interests, or demographics. This segmentation can aid in tailoring content, advertising, and engagement strategies to specific user segments.



$$J = \sum_{i=1}^{K} \sum_{j=1}^{n} ||x_j^{(i)} - \mu_i||^2$$

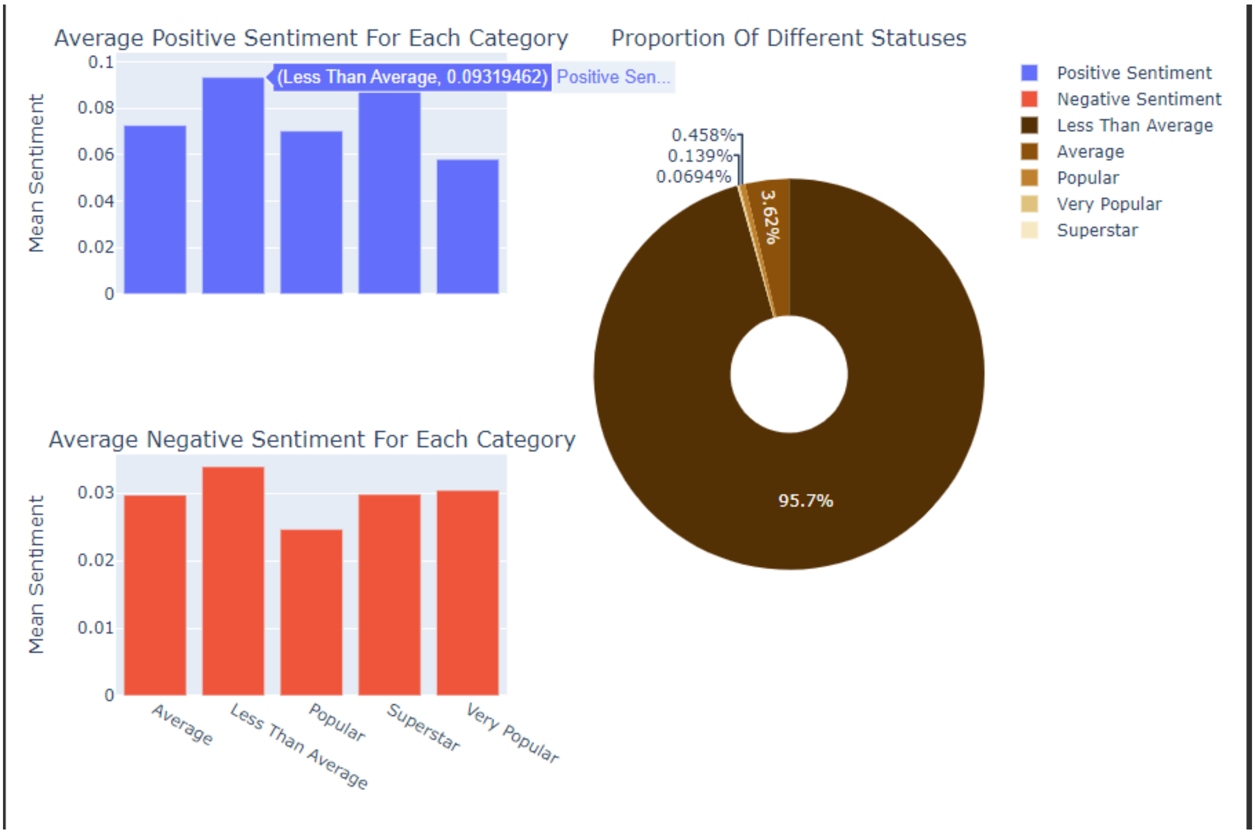
Where:

- K is the number of clusters.
- n is the number of data points.
- $x_j^{(i)}$ is the j-th data point in cluster i.
- μ_i is the centroid of cluster i.

How does the K-means algorithm work?

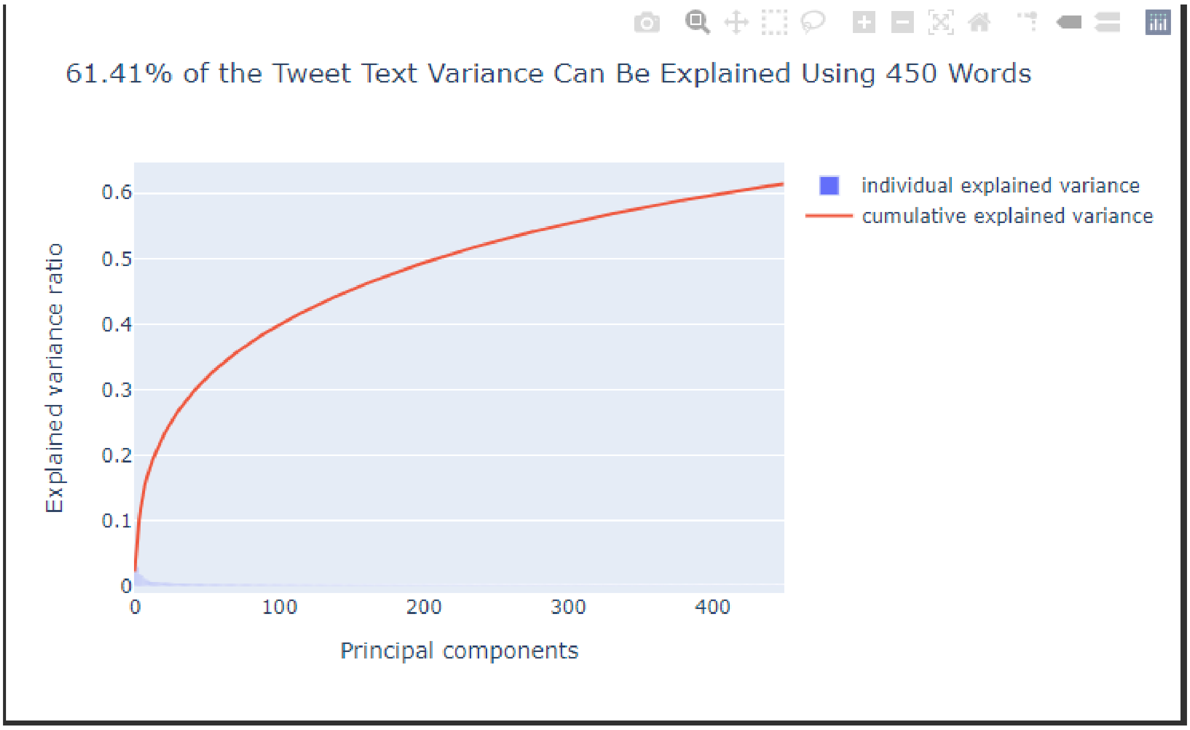
K-means clustering minimizes the sum of squared distances between each data point and its cluster's centroid. The formula is represented as J = ∑ (from i=1 to K) ∑ (from j=1 to n) ||x_j^(i) - μ_i||^2, where K is the number of clusters, n is the number of data points, x_j^(i) is a data point in cluster i, and μ_i is the centroid of cluster i. The algorithm iteratively assigns data points to the nearest centroids and updates centroids until convergence, creating K clusters that minimize the within-cluster variance.

5.8. Different Popularity Categories Average Sentiment Strength



We observe that the most popular Twitter users tend to be more extreme towards positivity and negativity in their tweets. And as to the regular, not popular users, they tend to be more positive on average in their tweet sentiment.

5.9. Text Decomposition Analysis



The total variance contribution of different words makes up our reduced dimension.

6. Conclusion

Sentiment analysis is a powerful and versatile tool that uncovers valuable insights from textual data across diverse domains. This document has delved into its intricacies, covering foundational principles, methodologies, data preprocessing, model application, and evaluation metrics. In our analysis of Pfizer Twitter data, sentiment analysis proves invaluable for deciphering the emotional undercurrents within user-generated content. From positive endorsements of the Pfizer BioNTech Covid vaccine to neutral information sharing and negative critiques, this analysis provides insights into public sentiment on a vital healthcare topic. Moreover, this analysis underscores the importance of harnessing sentiment analysis to not only comprehend public sentiment but also to inform strategies, shape policies, and better serve the needs of a connected world.

7.References

[1] Dataset: Pfizer Vaccine Tweets https://www.kaggle.com/datasets/gpreda/pfizer-vaccine-tweets

[2] Mongo DB Atlas Documentation https://www.mongodb.com/docs/atlas/

[3] Pyspark Tutorial https://sparkbyexamples.com/pyspark-tutorial/

[4] PyMongo 4.5.0 Documentation https://pymongo.readthedocs.io/en/stable/

[5] Tweet Sentiment Analysis [Preprocessing] https://www.kaggle.com/code/redwankarimsony/nlp-101-tweet-sentiment-analysis-preprocessing

[6] Text Blob Documentation https://textblob.readthedocs.io/en/dev/

[7] Count vectorization https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text/

[8] Logistic Regression https://www.geeksforgeeks.org/understanding-logistic-regression/

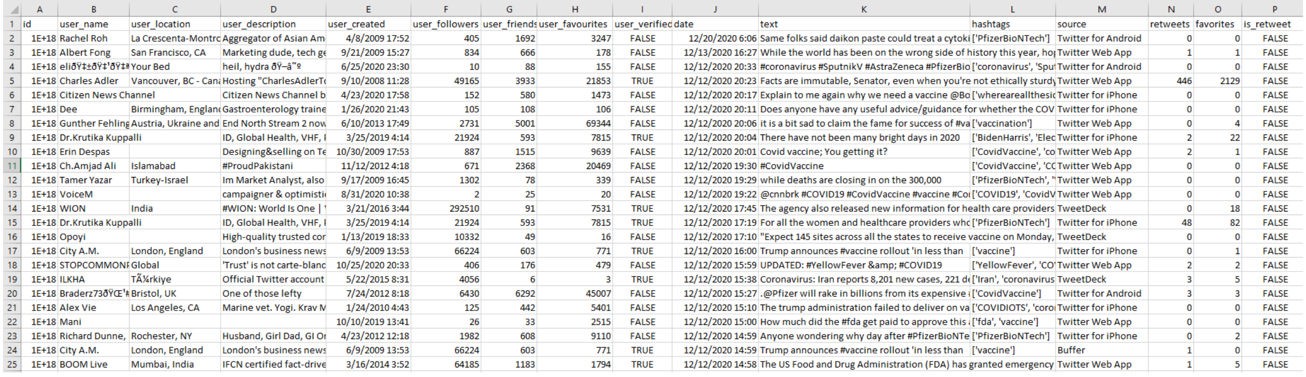
[9] Support Vector Machine https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/

[10] Hyperparameter Tuning using Grid Search https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/

[11] Confusion Matrix https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

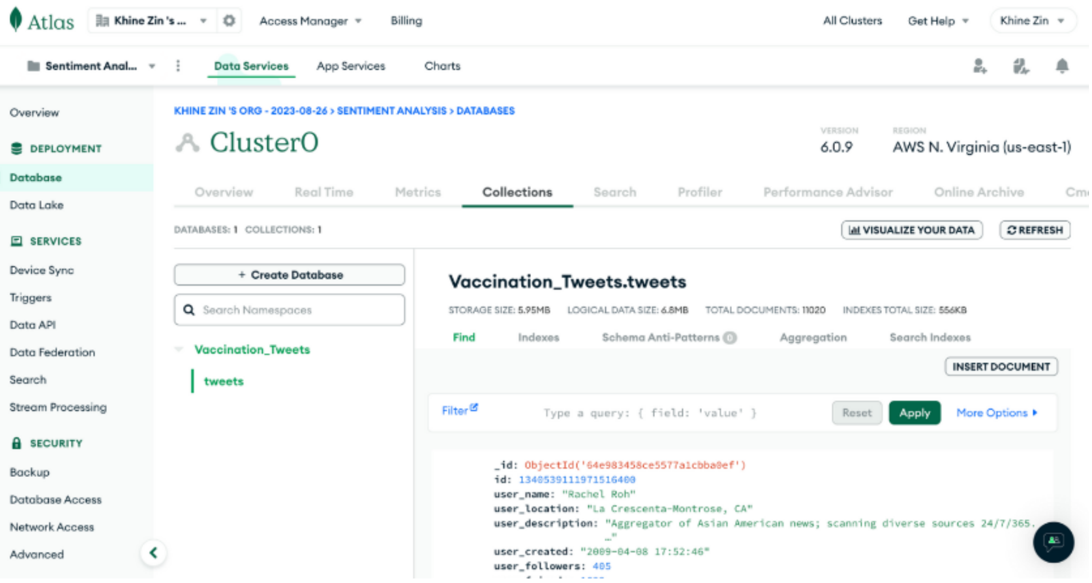
[12] K-Mean https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Appendix



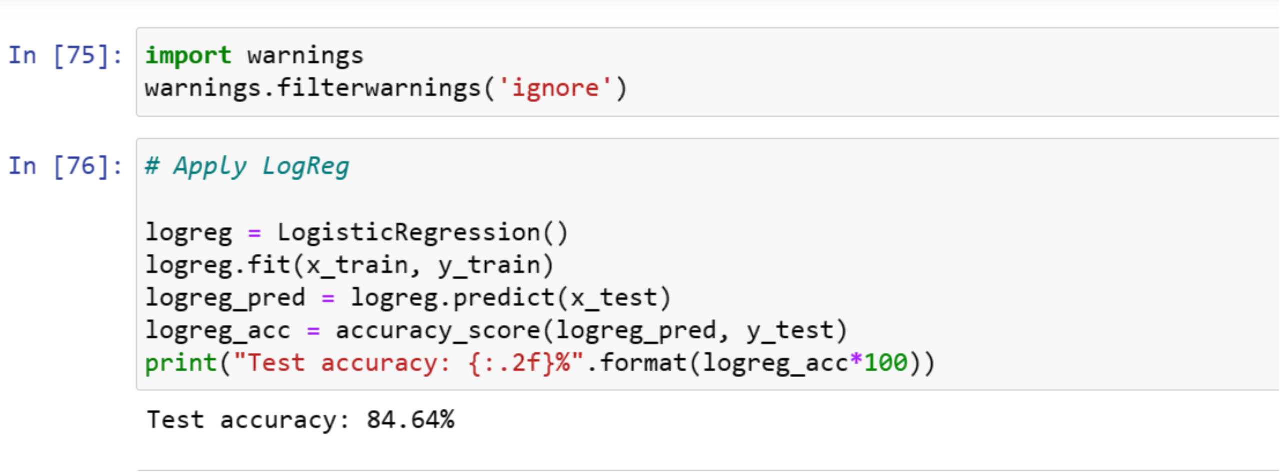
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	retweets	favorites	is_retweet
2	1E+18	Rachel Roh	La Crescenta-Montrc	Aggregator of Asian Am	4/8/2009 17:52	405	1692	3247	FALSE	12/20/2020 6:06	Same folks said daikon paste could treat a cytoki	['PfizerBioNTech']	Twitter for Android	0	0	FALSE
3	1E+18	Albert Fong	San Francisco, CA	Marketing dude, tech ge	9/21/2009 15:27	834	666	178	FALSE	12/13/2020 16:27	While the world has been on the wrong side of history this year, hoj		Twitter Web App	1	1	FALSE
4	1E+18	eliðŸ‡±ðŸ‡¹ðŸ‡ª	Your Bed	heil, hydra ðŸ–â˜º	6/25/2020 23:30	10	88	155	FALSE	12/12/2020 20:33	#coronavirus #SputnikV #AstraZeneca #PfizerBio	['coronavirus', 'Sput	Twitter for Android	0	0	FALSE
5	1E+18	Charles Adler	Vancouver, BC - Cana	Hosting "CharlesAdlerTc	9/10/2008 11:28	49165	3933	21853	TRUE	12/12/2020 20:23	Facts are immutable, Senator, even when you're not ethically sturdy		Twitter Web App	446	2129	FALSE
6	1E+18	Citizen News Channel		Citizen News Channel b	4/23/2020 17:58	152	580	1473	FALSE	12/12/2020 20:17	Explain to me again why we need a vaccine @Bo	['whereareallthesic	Twitter for iPhone	0	0	FALSE
7	1E+18	Dee	Birmingham, Englanc	Gastroenterology traine	1/26/2020 21:43	105	108	106	FALSE	12/12/2020 20:11	Does anyone have any useful advice/guidance for whether the COV		Twitter for iPhone	0	0	FALSE
8	1E+18	Gunther Fehling	Austria, Ukraine and	End North Stream 2 now	6/10/2013 17:49	2731	5001	69344	FALSE	12/12/2020 20:06	it is a bit sad to claim the fame for success of #va	['vaccination']	Twitter Web App	0	4	FALSE
9	1E+18	Dr.Krutika Kuppalli		ID, Global Health, VHF, I	3/25/2019 4:14	21924	593	7815	TRUE	12/12/2020 20:04	There have not been many bright days in 2020	['BidenHarris', 'Elec	Twitter for iPhone	2	22	FALSE
10	1E+18	Erin Despas		Designing&selling on Te	10/30/2009 17:53	887	1515	9639	FALSE	12/12/2020 20:01	Covid vaccine; You getting it?	['CovidVaccine', 'co	Twitter Web App	2	1	FALSE
11	1E+18	Ch.Amjad Ali	Islamabad	#ProudPakistani	11/12/2012 4:18	671	2368	20469	FALSE	12/12/2020 19:30	#CovidVaccine	['CovidVaccine', 'CC	Twitter Web App	0	0	FALSE
12	1E+18	Tamer Yazar	Turkey-Israel	Im Market Analyst, also	9/17/2009 16:45	1302	78	339	FALSE	12/12/2020 19:29	while deaths are closing in on the 300,000	['PfizerBioNTech', "	Twitter Web App	0	0	FALSE
13	1E+18	VoiceM		campaigner & optimisti	8/31/2020 10:38	2	25	20	FALSE	12/12/2020 19:22	@cnnbrk #COVID19 #CovidVaccine #vaccine #Coı	['COVID19', 'CovidV	Twitter Web App	0	0	FALSE
14	1E+18	WION	India	#WION: World Is One | '	3/21/2016 3:44	292510	91	7531	TRUE	12/12/2020 17:45	The agency also released new information for health care providers		TweetDeck	0	18	FALSE
15	1E+18	Dr.Krutika Kuppalli		ID, Global Health, VHF, I	3/25/2019 4:14	21924	593	7815	TRUE	12/12/2020 17:19	For all the women and healthcare providers whc	['PfizerBioNTech']	Twitter for iPhone	48	82	FALSE
16	1E+18	Opoyi		High-quality trusted cor	1/13/2019 18:33	10332	49	16	FALSE	12/12/2020 17:10	"Expect 145 sites across all the states to receive vaccine on Monday,		TweetDeck	0	0	FALSE
17	1E+18	City A.M.	London, England	London's business news	6/9/2009 13:53	66224	603	771	TRUE	12/12/2020 16:00	Trump announces #vaccine rollout 'in less than	['vaccine']	Twitter for iPhone	0	1	FALSE
18	1E+18	STOPCOMMONI	Global	'Trust' is not carte-blanc	10/25/2020 20:33	406	176	479	FALSE	12/12/2020 15:59	UPDATED: #YellowFever &amp; #COVID19	['YellowFever', 'CO'	Twitter Web App	2	2	FALSE
19	1E+18	ILKHA	TÃ¼rkiye	Official Twitter account	5/22/2015 8:31	4056	6	3	TRUE	12/12/2020 15:38	Coronavirus: Iran reports 8,201 new cases, 221 de	['Iran', 'coronavirus	TweetDeck	3	5	FALSE
20	1E+18	Braderz73ðŸŒ¹	Bristol, UK	One of those lefty	7/24/2012 8:18	6430	6292	45007	FALSE	12/12/2020 15:27	.@Pfizer will rake in billions from its expensive :	['CovidVaccine']	Twitter for Android	3	3	FALSE
21	1E+18	Alex Vie	Los Angeles, CA	Marine vet. Yogi. Krav M	1/24/2010 4:43	125	442	5401	FALSE	12/12/2020 15:10	The trump administration failed to deliver on va	['COVIDIOTS', 'coroı	Twitter for iPhone	0	0	FALSE
22	1E+18	Mani			10/10/2019 13:41	26	33	2515	FALSE	12/12/2020 15:00	How much did the #fda get paid to approve this :	['fda', 'vaccine']	Twitter Web App	0	0	FALSE
23	1E+18	Richard Dunne,	Rochester, NY	Husband, Girl Dad, GI Or	4/23/2012 12:18	1982	608	9110	FALSE	12/12/2020 14:59	Anyone wondering why day after #PfizerBioNTe	['PfizerBioNTech']	Twitter for iPhone	0	2	FALSE
24	1E+18	City A.M.	London, England	London's business news	6/9/2009 13:53	66224	603	771	TRUE	12/12/2020 14:59	Trump announces #vaccine rollout 'in less than	['vaccine']	Buffer	1	0	FALSE
25	1E+18	BOOM Live	Mumbai, India	IFCN certified fact-drive	3/16/2014 3:52	64185	1183	1794	TRUE	12/12/2020 14:58	The US Food and Drug Administration (FDA) has granted emergency		Twitter Web App	1	5	FALSE

Figure (1) CSV view of Pfizer Vaccine Tweets dataset



```
_id: ObjectId('64e983458ce5577a1cbba0ef')
id: 1340539111971516400
user_name: "Rachel Roh"
user_location: "La Crescenta-Montrose, CA"
user_description: "Aggregator of Asian American news; scanning diverse sources 24/7/365.
                  …"
user_created: "2009-04-08 17:52:46"
user_followers: 405
```

Figure (2) Mongo DB Atlas Database



In [75]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [76]:
```
# Apply LogReg

logreg = LogisticRegression()
logreg.fit(x_train, y_train)
logreg_pred = logreg.predict(x_test)
logreg_acc = accuracy_score(logreg_pred, y_test)
print("Test accuracy: {:.2f}%".format(logreg_acc*100))
```

```
Test accuracy: 84.64%
```

Figure (3) Python using Jupiter Notebook tool