COVID-19 Vaccines Related Tweets Sentiment Analysis and Relationship with Vaccination Levels

Fantastic Four:
Sharlene Chen
Jing Yang
Luqi Cai
Xueying Hu

Contents

1.Introduction

The prevalence of vaccines is an extremely important step in our battle with COVID, so we are curious about how society has been reacting to COVID vaccines. To do this, we found two datasets both from Kaggle, that could help us analyze the relationship between vaccination levels and people's sentiments towards COVID vaccines. We will perform sentiment analysis on these tweets, merge the two data sets based on country and time, and try to observe the relationship between sentiment and actual vaccination levels. As a result, we can help decipher what kind of sentiments could possibly help increase vaccination rates, and therefore help inform society on how to continuously improve vaccination rates.

2.Data Preprocessing

2.1 Dataset 1: Vaccination Tweers Cleaning

Our first dataset is vaccine-related tweets and user information ranging from December 2020 to November 2021. It consists of 200k+ records with 16 columns and we chose these 6 variables to analyze: user location, twitter date, text, retweets, and favorites. we dropped 29% of the records where the user locations are null.

Location Preprocessing since the user locations are not in the same format, we searched the location strings on the *Google Maps API* to extract the names of the country from the search results if they exist. For texts in non-English languages, we use pygtrans library to translate them into locations in English.

2.2 Dataset 1: Tweets Preprocessing

Text Preprocessing After cleaning up our first dataset column-wise, the next step was to clean each tweet. We referenced a paper on natural language processing and tweaked its method to match our needs to clean our tweets. The process is displayed on the right graph1.

Figure1: Flowchart for text preprocessing

Weighted Method We observed that some tweets have retweets and favorites. These two metrics could have similar emotional implications, but we thought retweets might be stronger compared to favorites. Thus, we decided to use retweets to calculate weight in sentiment analysis. However, we found that some tweets have more than several thousands of retweets. If we directly use the number of retweets as the weights, those tweets with the greater number of retweets will dominate the sentiment, which exaggerates the effects. To reduce the bias, we used $log(n+1+1)$ as the weight. n is the number of retweets, the first '1' represents the user who posted the tweet, and the second '1' is the number for avoiding the situation that if there is no retweet, then the weight would be log(1) = 0.

2.3 Dataset 2: Vaccination by Country and Time Dataset Cleaning

Our second dataset consists of 80k+ records with 16 attributes, and we used 6 attributes to conduct our analysis: Country, Date, Total number of people vaccinated, Daily vaccinations, Total number of people vaccinated per hundred, and Daily vaccinations per million. We converted the dates into a Date Time object and replaced the dates with the new format of (year, month) because we will be analyzing on a monthly level.

After basic cleaning, we then proceeded to separate the dataset into daily metrics and total metrics, to calculate the summation and group by easily. We first created a table that includes the daily vaccination and daily vaccinations per million. Next, we created a table that included all total vaccination metrics. In this table, we first used the two columns: people vaccinated, and people vaccinated per hundred. We then grouped the tables by country and year_month with the maximum.

Lastly, we combined the daily table and the total table by merging the indexes of the same country and same year_month. This will give us the metrics to see how different sentiments result in different daily and total vaccination levels.

2.4 Final Data Preprocessing: Merging Two Datasets

We performed an inner merge on the two datasets to gain comprehensive information on both vaccine and tweets text.

3.Sentiment Analysis

3.1 Simple Sentiment Analysis

To analyze the emotions within tweets, the first method we chose is simple sentiment analysis by Hu and Liu's sentiment analysis lexicon, in which words are coded as either positive or negative.

After we got the dictionary, we did the counting process. Since we considered the different weights for texts with different retweets, we added their own weights to the count instead of simple 1's. After that, we did two types of calculations. We repeated the whole process for each country each month to get the whole table. The formular are shown on the right and the result is *Table 1 in Appendix*.

positive_emotion_percentage = pos/(pos+neg)
negative_emotion_percentage = neg/(pos+neg)
positive_emotion = pos/number of all words
negative_emotion = neg/number of all words

Figure2: Formular for simple sentiments analysis

3.2 NRC Sentiment Analysis

For the emotion dictionary, we used NRC emotion lexicon wordleve v0.92, and used a for loop to extract the emotions in texts and multiply the emotion binary value by weight, so that we can get an emotion table for each country and month during the pandemic. The result is *Table 2 in Appendix*.

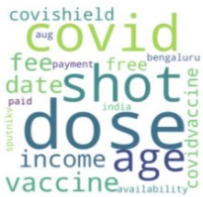
3.3 Vader Sentiment Analysis

Vader is a sentiment analysis tool that targets sentiments expressed in social media and is able to identify the logic within sentences, making it more advanced than simple and NRC sentiment analysis. We constructed the weight for each sentence and gathered them to country and month levels. And then we used Vader Sentiment to derive positive, negative, neutral, and compound emotions.The result is *Table 3 in Appendix*.

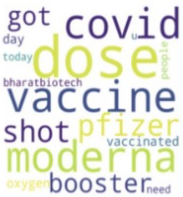
4.Results: EDA and Findings

4.1 Key Interesting Findings From EDA

Word clouds We visualized the 20 most frequent words in the top 5 countries in terms of the number of tweets. As shown in the word clouds below, in developing countries like India, the topic 'availability' and 'fee' are more popular; while for developed countries like the U.S. and Canada, the different kinds of vaccination are more frequent in the tweets. Additionally, we found out that across the timeline, 'age' is becoming more frequently used.

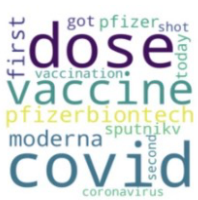


India



The U.S.

Canada



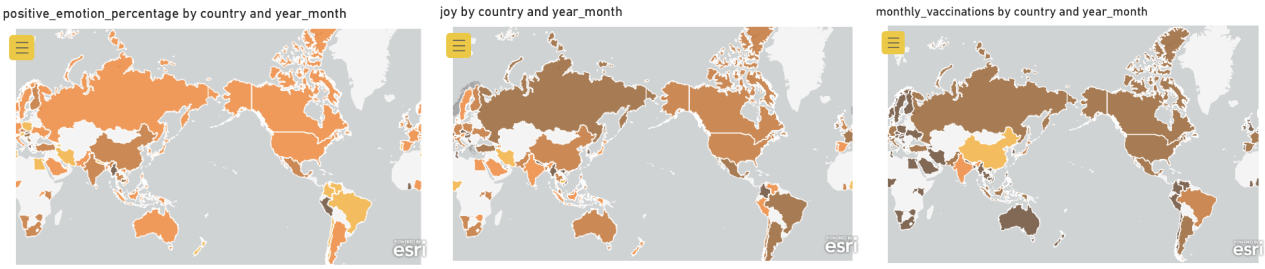
2020.12-2021.6

2021.7-2021.11

Heat Map We utilized Power BI to compare the relationship between emotions and monthly vaccination counts in each country. The heat maps display the whole world's emotion and vaccination progress, the bar charts compare the top 5 countries with the most tweets. From the graph below, we can see that there is no obvious relationship between monthly

vaccination counts and positive emotion percentage, but geographical relationships do exist. *All figures (figure 4,5) are in the appendix*.

Additionally, we also observed a positive relationship between monthly vaccination number surge within one month and positive emotion percentage, as we can see in the graph below. We also discovered the relationship between joy emotion and monthly vaccination number and will further investigate how each emotion relates to the monthly vaccination and verify this using Machine Learning models.



4.2 Model Comparison and Findings

Linear Regression After obtaining all sentiment scores, we created several regression models to measure how each sentiment is correlated to vaccination levels. The independent variables are the emotions, and the dependent variables are monthly vaccination levels in percentage. The dataset was also split into a train and a test set. All regression results are displayed in *Table7 in Appendix*.

Linear Regression for simple sentiment analysis The first regression model is based on the simple sentiment analysis. Positive emotion has a significant p-value below 0.05, and has a positive correlation to vaccination levels, meaning that positive emotion does contribute significantly to higher vaccination rates.

Linear Regression for NRC sentiment analysis The second regression model fits NRC's 8 sentiments to vaccination levels. We can see that joy and disgust are the most significant emotions. Joy has positive coefficients – the more joy, the more people accept vaccinations. This also aligns with the first simple model that we fitted. Another emotion – disgust – may be a bit less intuitive. However, if we consider the impacts of policies that could generate disgust, the more disgust may mean more restrictions and policies put in place and thus lead to higher vaccination rates. But since the overall R-squared score is low, this suggests that many other factors are not captured in emotions that also impact vaccination levels.

Linear Regression for NRC sentiment analysis The last linear regression was fitted with Vader Sentiments. However, Vader Sentiment results don't give statistically significant variables, the most significant only has a p-value of 0.11, this could be explained by high multicollinearity that is likely to be present since the scores depend on each other. Thus, more data about policies and emotions or other methods of fitting are needed to confirm the relationship between sentiments and vaccination levels. Comparison results can be found in *Table 8 in Appendix*

Random Forest The team used the NRC sentiment and Vader sentiment scores to fit the data with the random forest algorithm. GridCV was used to hyper-tune the parameters of the model. The following results are the best hyper-tuned parameters for the best performing model (fitted from Vader sentiments): max depth: 6; min sample leaf: 50; min sample split: 50; N Estimators: 6.

Here, the best RMSE is about 46.48 from Vader Sentiments, though it varies a bit when run multiple times. The important features of NRC sentiments include disgust, fear, and joy which match results from linear regression. The important feature of Vader Sentiments is negative emotions. Neutral ranks second, which matches with the result of linear regression. Results can be found in *Table 8 in Appendix*

k-Nearest Neighbors NRC and Vader sentiments were also fitted with the k-Nearest Neighbors' algorithm (specifically k-nearest regressors). The best k in kNN for NRC sentiments is 201, and the RMSE is 49.06, while the best k for the Vader Sentiment kNN algorithm is 128, and the RMSE is 46.39. Though the results are better than the previous two models, kNN does not allow us to figure out which emotions are the most significant. The below diagrams show the loss based on different ks.

Multi-Layer Perceptron Regressor We've also fitted a Multi-Layer Perceptron model on the Vader Sentiment scores. The best model from GridCV has the solver: adam, 4 hidden layers, iteration of 1000, tanh activation function, and an adaptive learning rate. The resulting RMSE is 46.69, which is higher than linear regression and Random forest, but lower than kNN.

Overall, Vader Sentiments gives the best results in all 4 methods. Linear regression can best tell us what emotions are more significant, but kNN performs best in terms of mean squared error for predicting vaccination rates. R-squared in general is quite low, showing that there are many other factors impacting the dependent variable (vaccinated rates). The most significant emotions are positive emotions for 2 emotions, or within the 8 emotions context, the most significant ones are disgust and joy. More evidence is needed to prove the relationship since results don't all align well..

	Linear Regression	Random Forest	kNN	MLP Regressor
Best Setup	Vader Sentiments Regression	Vader Sentiments Random Forest max depth: 6; min sample leaf: 50 min sample split: 50; n estimators: 6	Vader Sentiments k=128	solver: Adam hidden layer: 4 max iter: 1000 activation: tanh learning rate: adaptive
Best Root Mean Squared Error	46.54	46.48	46.39	46.64
Most Significant Variables	2 sentiment: Positive 8 sentiment: Disgust and Joy	8 sentiment: Disgust, Fear, Joy Vader: Negative	N/A	N/A

Table 1: Final Comparison Between All Four Models

5.Conclusions

Though from simple sentiment analysis and NRC analysis, we have seen that positive emotions are significant and have positive correlations with vaccination rates, the model that has the lowest root mean squared error (the models constructed with Vader Sentiment scores) doesn't align with this result. Thus, we can't conclude that sentiments would impact vaccination rates, we believe that further text data on policies within a certain period/country is needed to prove the relationship. Furthermore, we could also include topic analyses (LDA) to perhaps further identify certain topics that may have appeared at different time points in different countries to identify how different topics related to vaccination rates.

6.Appendix

positive_emotion_percentage	negative_emotion_percentage	positive_emotion	negative_emotion
54.536643	45.463357	2.749019	2.291664
69.897000	30.103000	1.477365	0.636266
48.960058	51.039942	3.139462	3.272830
53.724357	46.275643	8.166725	7.034434
0.000000	100.000000	0.000000	5.882353

Table 1: The result for simple sentiment analysis

	disgust	surprise	fear	positive	joy	anger	negative	sadness	anticipation	trust
0	0.000000	0.000000	0.000000	25.000000	0.000000	0.000000	0.000000	0.000000	12.500000	0.000000
1	0.000000	0.000000	0.000000	13.513514	0.000000	0.000000	0.000000	0.000000	0.000000	5.405405
2	3.144267	1.886560	7.914098	17.975754	0.628853	4.401974	8.542951	4.401974	8.542951	5.659681
3	3.508772	0.000000	5.263158	17.543860	0.000000	3.508772	5.263158	5.263158	5.263158	8.771930
4	0.000000	0.000000	0.000000	22.222222	0.000000	0.000000	11.111111	0.000000	0.000000	0.000000

Table 2: The result for NRC sentiment analysis

Vader pos	Vader neg	Vader neu	Vader com
0.000000	0.000000	1.000000	0.000000
0.119500	0.071000	0.809500	0.148000
0.043696	0.017581	0.842476	0.060754
0.036143	0.068214	0.752786	-0.064843
0.174000	0.000000	0.826000	0.378600

Table 3: The result for vader sentiment analysis

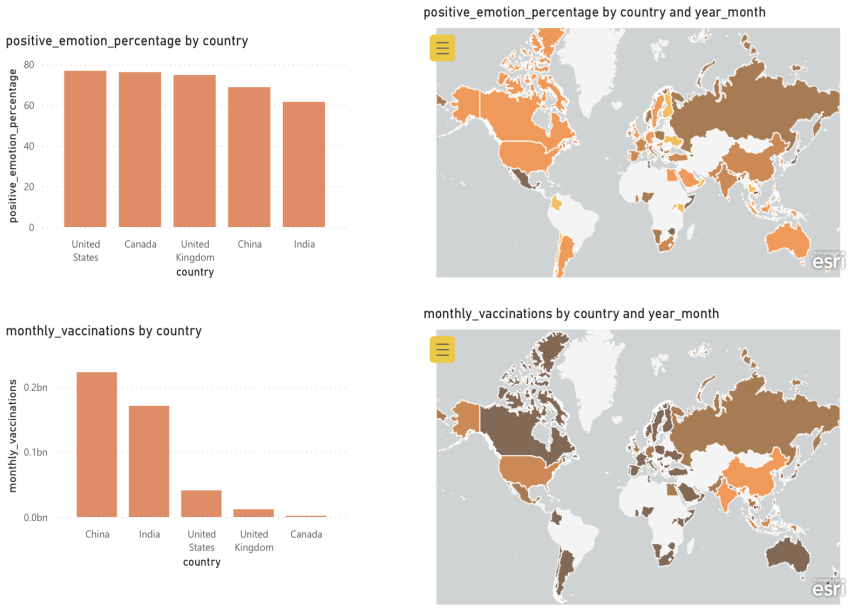


Figure 4: Heat map comparing the relationship between positive emotion and monthly vaccination counts

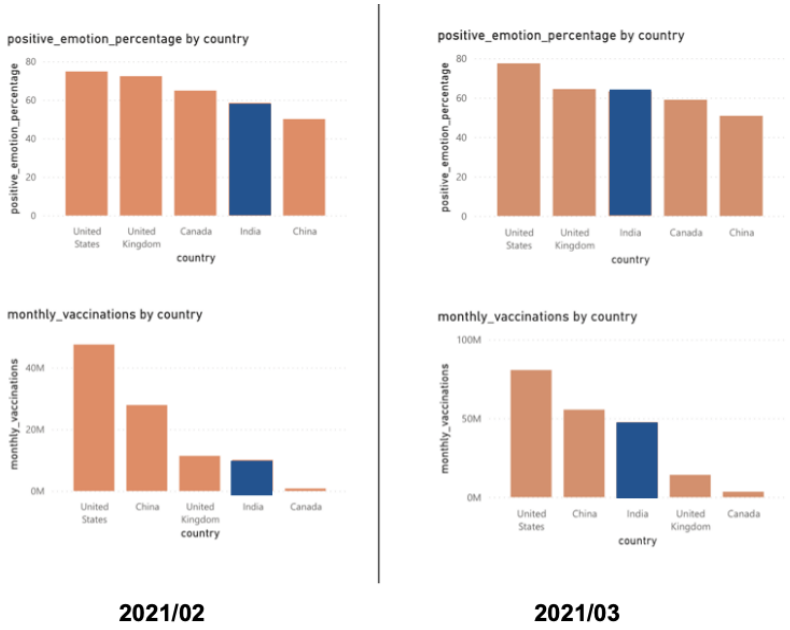


Figure 5: Bar charts showing the relationship between monthly vaccination surge and increase in positive emotion

```
test
```

	disgust	fear	joy	surprise	anger	trust	sadness	anticipation	daily_vaccinations_per
521	1.787902	3.494536	3.460807	2.275512	1.869171	6.860398	2.194244	3.833419	2.8028
727	1.597956	6.261263	3.937532	1.376885	2.753770	4.470184	2.974841	5.432875	14.3303
477	0.000000	0.000000	2.777778	2.777778	0.000000	5.555556	0.000000	2.777778	2.0914
730	0.000000	3.260078	5.843628	3.260078	0.000000	9.628461	3.260078	5.843628	41.1664
461	4.259290	8.578853	3.623150	2.634082	3.732474	5.859973	6.496114	5.288397	24.1894
...	...	...	...	...	...	...	...	...	...
709	1.351139	4.004209	4.350719	4.161225	3.146730	6.241413	3.978454	5.169302	7.1467
633	1.837513	3.328061	3.185917	2.332091	2.796894	7.687170	3.328061	6.091233	13.3287
162	1.826321	4.565802	1.826321	1.826321	3.652642	1.826321	3.652642	3.273646	20.1825
126	0.443189	3.634008	5.450880	2.291256	0.886379	9.528077	1.687380	9.787327	24.2127
244	0.000000	4.866793	0.000000	0.000000	0.000000	2.984061	4.866793	4.866793	1.1392

Table 6: Example Test Set

Dep. Variable:	daily_vaccinations_per	**R-squared:**	0.043
Model:	OLS	**Adj. R-squared:**	0.040
Method:	Least Squares	**F-statistic:**	16.53
Date:	Fri, 22 Apr 2022	**Prob (F-statistic):**	9.45e-08
Time:	15:46:28	**Log-Likelihood:**	-2718.7
No. Observations:	741	**AIC:**	5443.
Df Residuals:	738	**BIC:**	5457.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.8091	0.693	11.271	0.000	6.449	9.169
positive_emotion	0.3211	0.059	5.438	0.000	0.205	0.437
negative_emotion	0.1404	0.109	1.285	0.199	-0.074	0.355

Omnibus:	82.344	**Durbin-Watson:**	0.633
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	108.402
Skew:	0.928	**Prob(JB):**	2.89e-24
Kurtosis:	3.255	**Cond. No.**	21.7

Dep. Variable:	daily_vaccinations_per	**R-squared:**	0.030
Model:	OLS	**Adj. R-squared:**	0.019
Method:	Least Squares	**F-statistic:**	2.809
Date:	Mon, 25 Apr 2022	**Prob (F-statistic):**	0.00451
Time:	21:51:20	**Log-Likelihood:**	-2723.7
No. Observations:	741	**AIC:**	5465.
Df Residuals:	732	**BIC:**	5507.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.4806	0.777	13.494	0.000	8.956	12.005
disgust	0.6124	0.210	2.921	0.004	0.201	1.024
fear	0.1752	0.190	0.922	0.357	-0.198	0.548
joy	0.6315	0.195	3.238	0.001	0.249	1.014
surprise	-0.1087	0.200	-0.544	0.586	-0.501	0.283
anger	-0.0082	0.228	-0.036	0.971	-0.455	0.439
trust	-0.2016	0.125	-1.614	0.107	-0.447	0.044
sadness	-0.1101	0.221	-0.498	0.618	-0.544	0.324
anticipation	-0.2063	0.133	-1.551	0.121	-0.467	0.055

Omnibus:	77.459	**Durbin-Watson:**	0.615
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	100.766
Skew:	0.900	**Prob(JB):**	1.32e-22
Kurtosis:	3.152	**Cond. No.**	23.5

Dep. Variable:	monthly_vaccinations_per	**R-squared:**	0.004
Model:	OLS	**Adj. R-squared:**	-0.001
Method:	Least Squares	**F-statistic:**	0.8386
Date:	Fri, 29 Apr 2022	**Prob (F-statistic):**	0.501
Time:	15:09:19	**Log-Likelihood:**	-3007.6
No. Observations:	812	**AIC:**	6025.
Df Residuals:	807	**BIC:**	6049.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	14.4921	2.108	6.874	0.000	10.354	18.630
vader_pos	-9.6429	22.553	-0.428	0.669	-53.913	34.627
vader_neg	-1.4961	26.900	-0.056	0.956	-54.299	51.307
vader_neu	-4.1600	2.631	-1.581	0.114	-9.324	1.004
vader_com	4.0555	9.345	0.434	0.664	-14.287	22.398

Omnibus:	84.546	**Durbin-Watson:**	0.585
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	110.788
Skew:	0.904	**Prob(JB):**	8.77e-25
Kurtosis:	3.048	**Cond. No.**	124.

Table 7: Regression Model Results for 3 different sets of sentiments (Left: Simple, Center: NRC, Right: Vader)

	Simple Sentiments	NRC Sentiments	Vader Sentiments
Root Mean Squared Error	48.81	53.64	46.54
Most significant variables	Positive	Disgust, Fear, Joy	Vader Neutral

Table 8: Comparison of 3 Regression Results

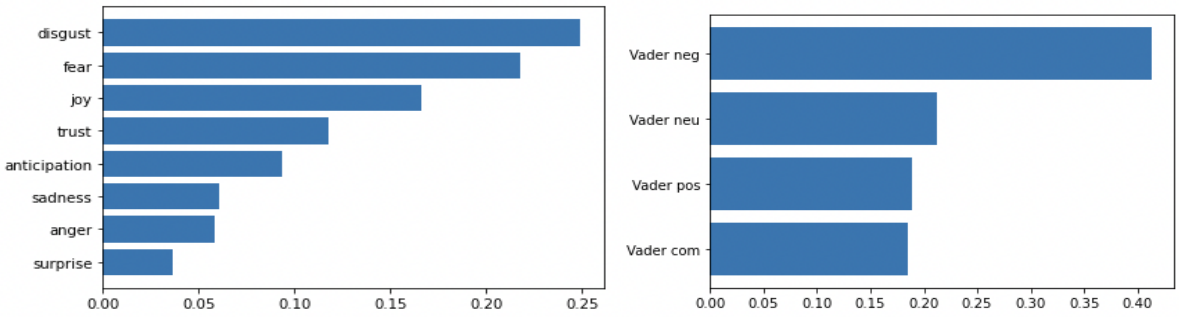


Figure 9: Comparison of Important Features for 2 Different Sentiment Sets

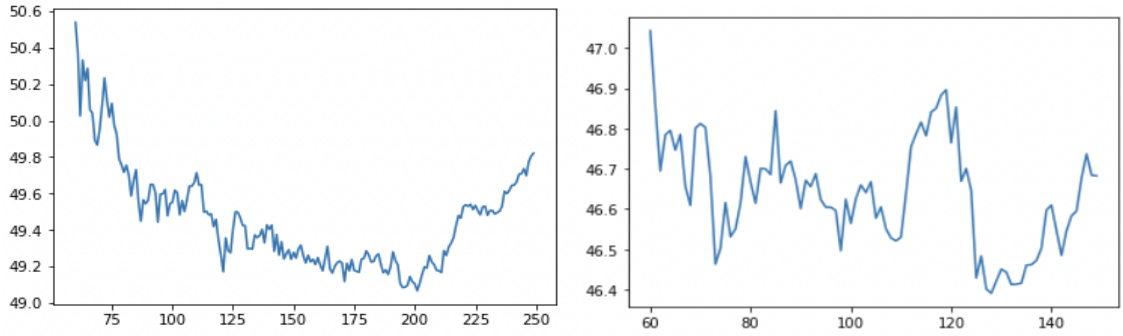


Figure 10 : Comparison of kNN fitting for 2 Different Sentiment Sets

7.Reference

Batra, R., Imran, A. S., Kastrati, Z., Ghafoor, A., Daudpota, S. M., & Shaikh, S. (2021). Evaluating Polarity Trend Amidst the Coronavirus Crisis in Peoples' Attitudes toward the Vaccination Drive. *Sustainability*, *13*(10), 5344.