Sentiment and Time series Analysis for Pfizer & BioNTech COVID-19 Vaccine

Aabha Desai

05/06/2021

Executive Summary

COVID-19, is an ongoing global pandemic of coronavirus disease 2019 caused by severe acute respiratory syndrome coronavirirus 2 (SARS CoV-2). The virus was first identified in Dec 2019 in Wuhan, china and on 11th of Mar 2020 WHO declared it a pandemic. With more than 147 million confirmed cases and more than 3.12 million deaths, this has been one of the deadliest pandemics in history.

Various COVID-19 vaccines have been developed to provide acquired immunity against this virus and as of today, 13 vaccines have been authorized by at least one national regulatory authority for public use all over the world. CDC has authorized and recommended 3 vaccines in the United States to prevent COVID-19:

1. Pfizer-BioNTech, 2. Moderna and 3. Johnson & Johnson/Janssen.

On December 11,2020, the U.S. Food and Drug Administration issued the first emergency use authorization (EUA) that allowed the Pfizer-BioNTech COVID-19 Vaccine to be distributed in the U.S. There has been a fair amount of concern regarding this vaccine and it will be critical for public health authorities to be able to effectively counter negative public attitude towards the vaccine.

In this report, sentiment and time series analysis has been used to answer three important questions regarding Pfizer-BioNtech COVID-19 vaccine:

1. Is willingness to be vaccinated increasing or decreasing?
2. What are some of the reasons for public hesitancy about getting vaccinated?
3. What motivates people most to get vaccinated?

From this analysis, we can conclude that, people are generally accepting towards Pfizer-BioNTech COVID-19 vaccine and are willing to get vaccinated. People having negative attitude towards the vaccine are more concerned about the side effects, it's safety and allergies. People showcasing positive sentiment seem to believe in scientific studies carried out to develop the vaccine, approvals given and efficacy of the vaccine against COVID-19. This analysis would help public health authorities design and plan vaccine campaigns to increase public inoculation by focusing their ads and campaigns on concerns that people have.

A. Introduction

Wearing a mask, maintaining 6 feet distance and washing your hands often, do help prevent getting infected, but the best way to overcome this pandemic is to get immunity against the virus. Vaccines teach our immune systems how to recognize and fight the virus. However, there has been a concern that the development of this COVID-19 vaccine was too rushed and was done under political pressure. Some people still believe that the risks of taking the vaccine outweighs the benefits. It is critical for public health authorities to be able to counter negative attitude towards the vaccine. The first step, however, is to understand the public opinion about the vaccine.

Social media provides a great deal of data that can be used for analyzing the overall public sentiment about any product. Since Pfizer-BioNTech COVID-19 Vaccine was the first one to start its distribution, ample data is available for its analysis. Here, tweets about the vaccine are used in assessing public opinions and acceptance of the Pfizer-BioNtech vaccine. However, this analysis can be extended to other vaccines too.

To help answer the questions mentioned in the executive summary, the following hypotheses have been analyzed:

1. The dominant sentiment about the vaccine is positive.
2. The daily average sentiment follows an upward trend.
3. The sentiment series is stationary and can be used to forecast average sentiment score in the future.
4. There is a strong correlation between the average sentiment score and the percentage of people vaccinated.
5. There is a strong correlation between the average sentiment score and daily number of tweets.

Now that the goal of Sentiment and Time series Analysis for Pfizer & BioNTech COVID-19 Vaccine has been introduced, it is important to understand the datasets utilized. The datasets with data descriptions have been introduced in the next section.

B. Data Description

Two datasets have been used for this analysis.

1. Pfizer & BioNTech Vaccine Tweets - This dataset includes tweets about Pfizer-BionTech Vaccines from 12 Dec 2020 to 15 Mar 2021. It provides 6818 records/observations with 16 variables such as id, user_name, user_description, user_followers, date, text, hashtags, retweets etc.

2. COVID-19 World Vaccination Progress - This dataset includes the Daily and Total Vaccination data for COVID-19 from 19 Dec 2020 to 15 Mar 2021, in the World.It provides 6517 records/observations with 15 variables such as country, date, vaccines, total_vaccinations, people_vaccinated, people_fully_vaccinated, daily_vaccinations etc.

To help with the analysis, some features have been added to both the datasets.

C. Preprocessing the Data

Loading the First Dataset - Pfizer & BioNTech Vaccine Tweets

As part of the initial step of preprocessing, the first dataset having Pfizer-BioNTech COVID-19 vaccine Tweets was imported and its data formatting was checked. In this dataset, the "date" and "user_created" variables were incorrectly classified as chr i.e. character string. These formats were corrected and the dataset was converted to a dataframe.

To help with the analysis, seven more columns were added to the original data frame. The "user_verified" column had boolean values, so those were changed to chr strings telling us whether the user was verified for each record. The blue verified badge on Twitter lets people know that an account of public interest is authentic. To receive the blue badge, your account must be authentic, notable (partnership or direct outreach with government,companies,brands, organizations,news organizations and journalist,activists,entertainment influential individuals), and active(must be active with a record of adherence to the Twitter Rules).

The column "account_age" tells us how old the user account is as of today. The column "total_engagement" is the sum of the number of retweets and favorites (likes for that tweet). The column "tweet_length" gives us the length of the tweet string which is further classified as 'short' or 'long' based on the string length. The text content of a tweet can contain a maximum of 280 characters. Based on this, Tweets having string length more than 140 characters were classified as 'long' and smaller than 140 characters were classified as 'short'. The column "hash" gives the number of hashtags in the tweet text. The column "mentions" gives the number of mentions in the tweet text. The column "Media" tells us if the tweet has any media such as photos or videos. The column "account_class" has categories like Weak(user_followers <100), Normal (100<user_followers<1000), Strong (1000<user_followers<10000) and Influencer (10000<user_followers) based on the number of followers the user has.

Loading the Second Dataset - COVID-19 World Vaccination Progress

Now the second dataset having COVID-19 World Vaccination Progress was imported and its data formatting was checked. In this dataset, the "date" variable was incorrectly classified as chr i.e. character string which was corrected and the dataset was converted to a dataframe.

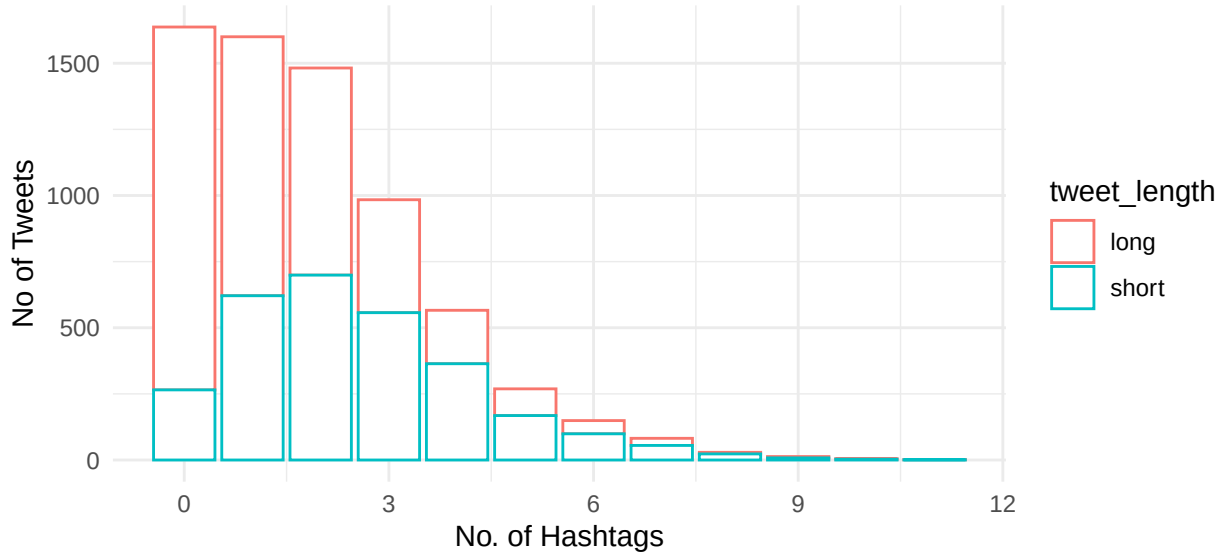
Some variables/columns not used for the analysis have been removed and the final dataframe consists of 9 variables.

D. Exploratory Data Analysis :

Before going into the testing phase, a preliminary analysis was conducted to identify the main characteristics of the datasets. This analysis was done to further understand the data and determine whether this data was adequate for testing the hypotheses.

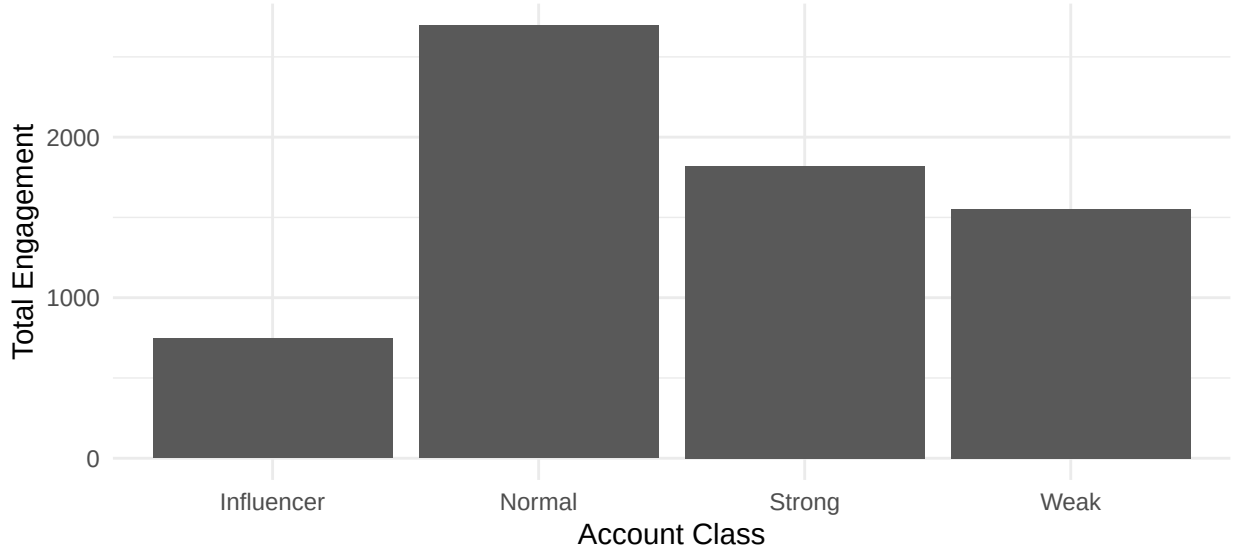
Pfizer - BioNTech Vaccine Tweets Dataset EDA

The barplot below shows the number of hashtags against the number of tweets having long and short tweet lengths.



As can be seen the number of Long length tweets decrease with increasing number of hashtags. However, the number of short length tweets increased with the number of hashtags in the beginning and then gradually decline. The maximum number of long length tweets had no hashtags, whereas the maximum number of short length tweets had 2 hashtags. Altogether, around 5260 tweets were long text tweets and 1558 were short length tweets.

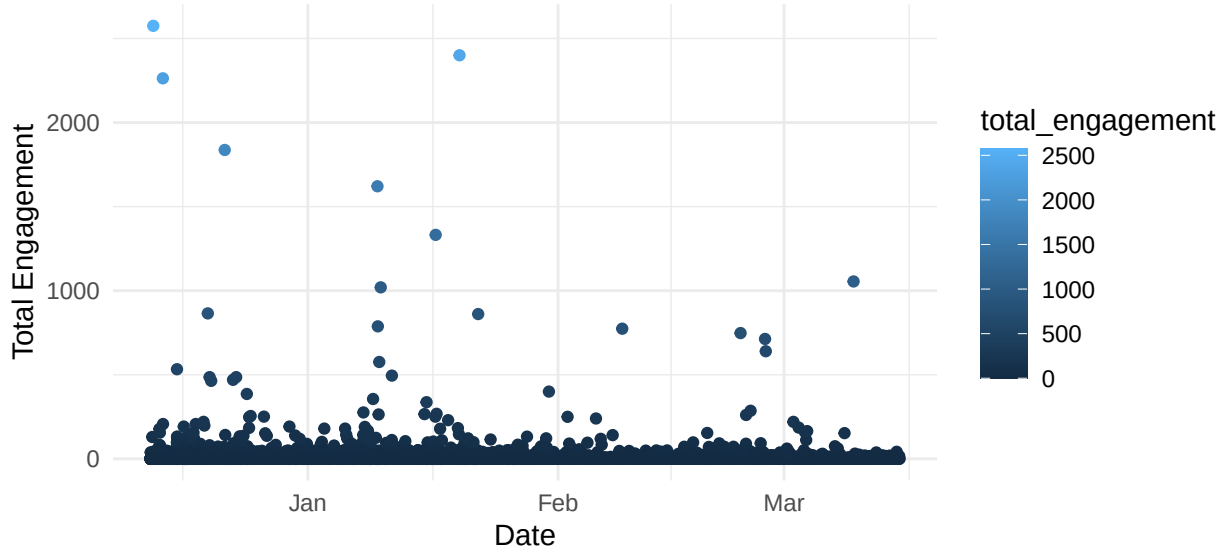
Next, we look at the Engagement by account class.



D. Exploratory Data Analysis :

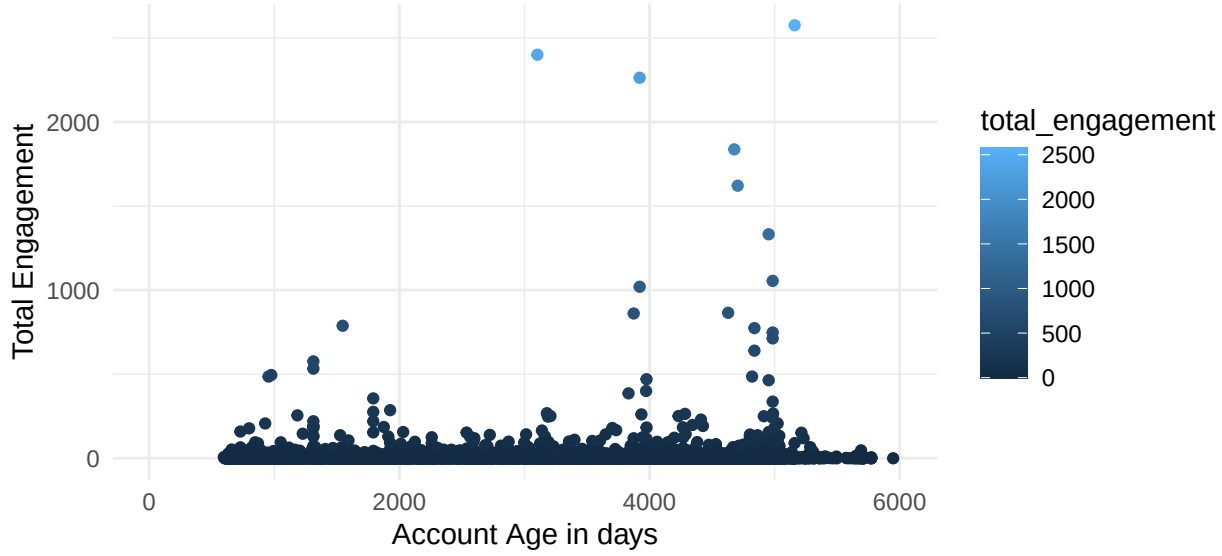
Surprisingly, above plot showed that the influencers (having followers > 10,000) had the least engagement and normal accounts having followers between 100 to 1000, had the most engagement.

In the plot below, the relation between total engagement with time was explored.



It was observed that the Total Engagement was maximum during Dec 2020 and quite low at the beginning of Jan 2021. We also saw some spikes in between.

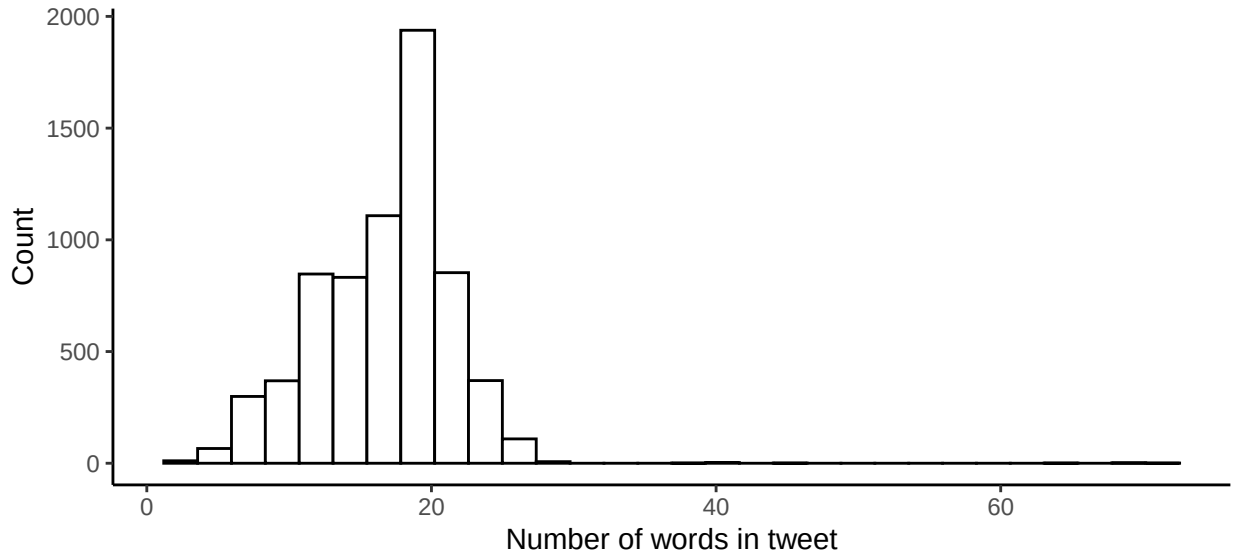
Next, we checked if the total engagement depends on the account age.



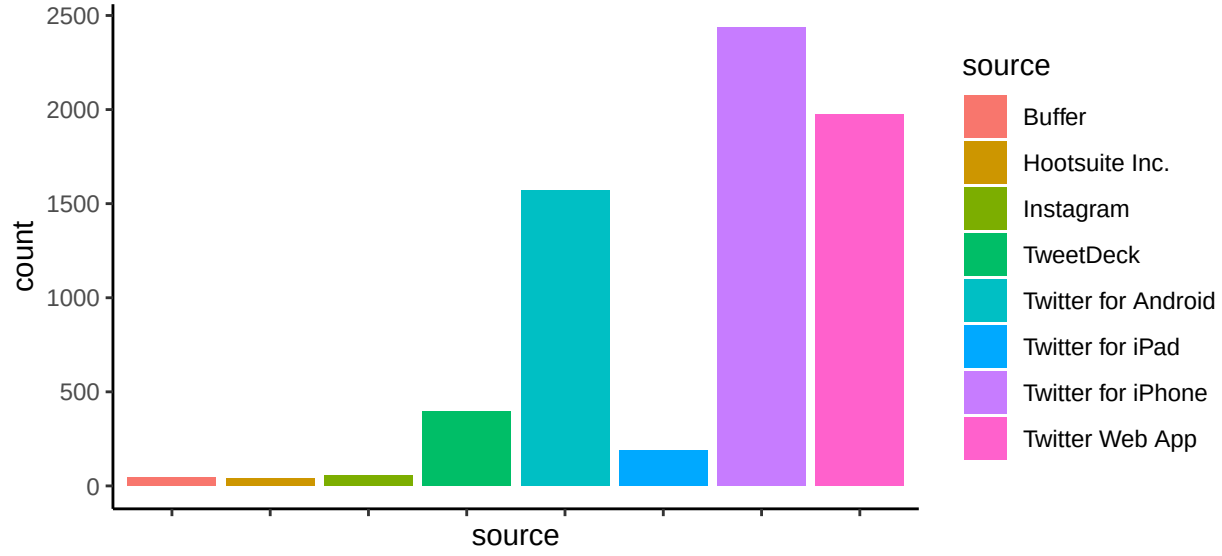
Overall the Total engagement did not show much change with increasing Account Age. We saw some increase in the total engagement between 4000 - 5000 day old accounts.

We also looked into the distribution of words.

D. Exploratory Data Analysis :



The count of tweets increased with the number of words in a tweet till it reached its maximum at 18 words per tweet. It decreased sharply after that, falling to zero for tweets with more than 30 words.

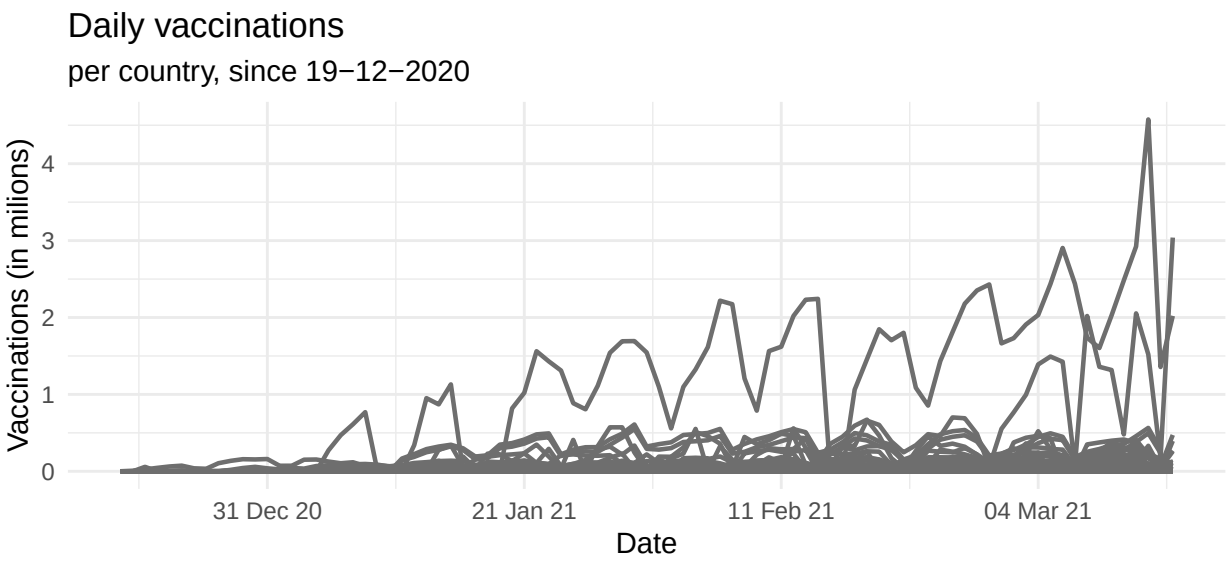


The dataset has 43 unique sources of platforms used to tweet and the number of tweets from "Twitter from iphone" were maximum and least for Buffer (Buffer is a Twitter app that allows you to add tweets to be sent on a schedule).

COVID-19 World Vaccination Progress Dataset EDA

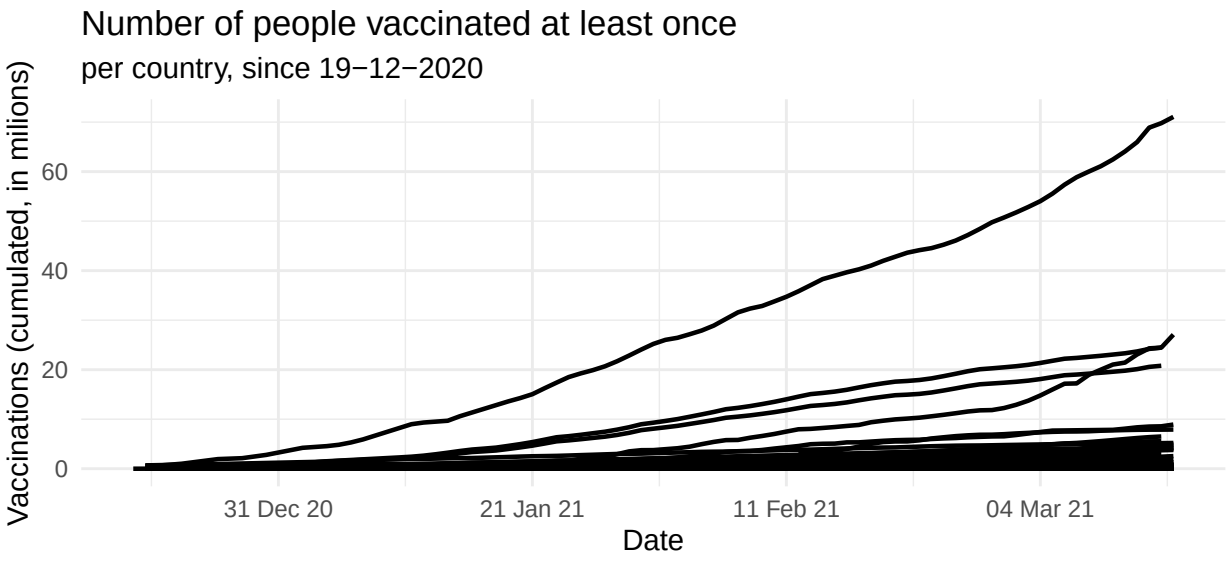
The following plot shows Daily Vaccinations in all the countries since 19 Dec 2020.

D. Exploratory Data Analysis :



Each line represents a country. Overall daily vaccinations increased from Dec 2020. Some countries showed more ups and downs than others.

Next we checked the number of people vaccinated atleast once, in all the countries since 19 Dec 2020.



Like the daily vaccination plot, the number of people vaccinated at least once across all the countries, increased from Dec 2020. Some countries showed better trend than others.

E. Empirical Analysis

Tweets NLP - Text Mining

Text mining is used to convert unstructured text from tweets into a structured text.Strings were converted to lower case and things unrelated to the sentiment analysis like URLs, spaces and special characters were removed to get a clean filtered text. An example of this is shown below:

```
## [1] "Before: Vaccine!! Anyone?? #covid #Pfizervaccine #PfizerBioNTech https://t.co/b9ZKwnlIkX"
```

```
## [1] "After: vaccine anyone covid pfizervaccine pfizerbiontech"
```

Vader Sentiment Analysis

There are many packages available for generating a sentiment score. The Vader package has been used for the present analysis. Vader stands for Valence Aware Dictionary and Sentiment Reasoning. It is used for textual sentiment analysis. It is sensitive to both polarity (positive/negative) and intensity (strength) of an emotion, which makes it suitable for the analysis.

How does this work?

This is based on lexicons of sentiment related words. Each of the words in the lexicon is rated as to whether it is positive or negative, and in many cases, how positive or negative it is.

What is polarity?

The key aspect of sentiment analysis is to analyze a body of text for understanding the opinion expressed by it. Here, we quantify this sentiment with a positive or negative value, called polarity. The overall sentiment is often inferred as positive, neutral or negative from the sign of the polarity score.

What is compound score?

It calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).

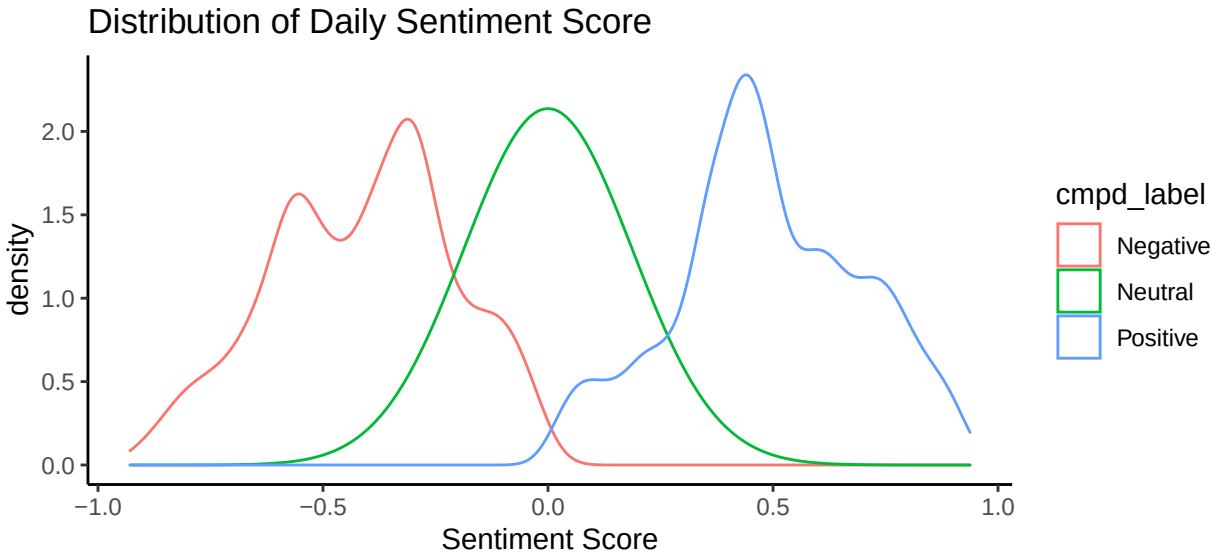
After the compound scores were computed, a label called “cmpd_label” was created for all the records. This variable assigned “Positive” label if the compound score was greater than 0, “Negative” if the compound score was less than 0 and “Neutral” if the compound score was equal to 0. This labeling helps us differentiate between sentiments easily.

The questions mentioned in the executive summary were answered with the help of the following hypotheses testing:

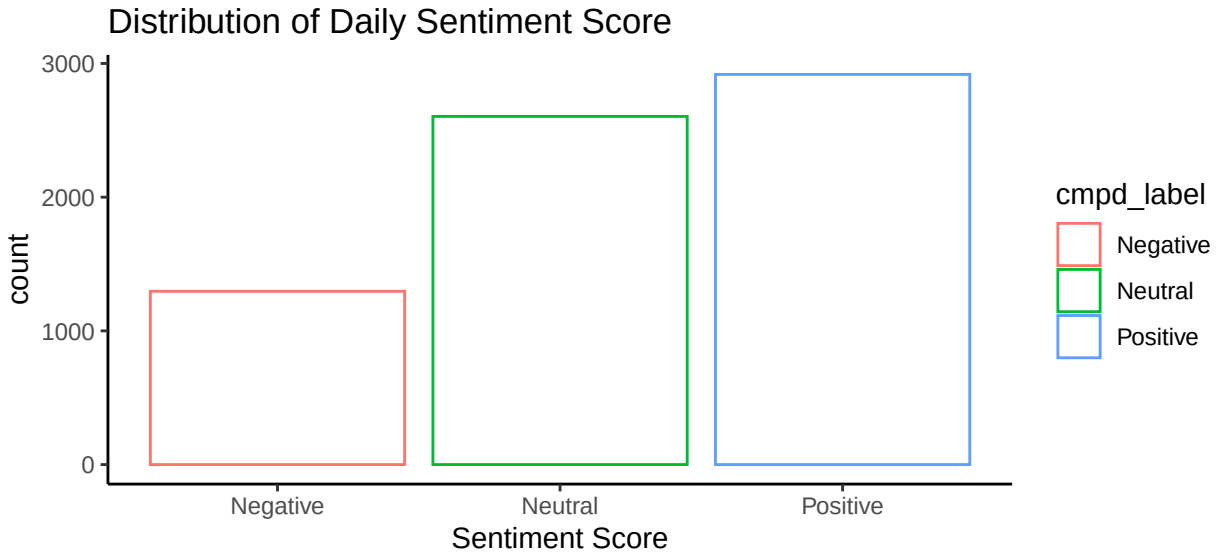
1. Is willingness to be vaccinated increasing or decreasing?

Hypothesis 1: The dominant sentiment about the vaccine is Positive.

To test this hypothesis, the density and box plots of daily sentiments were analyzed.



The density plot showed that the distributions of the sentiments was approximately normal. The Neutral sentiment had a standard normal distribution. Overall, the peak of Positive sentiment was more than that of Negative or Neutral.

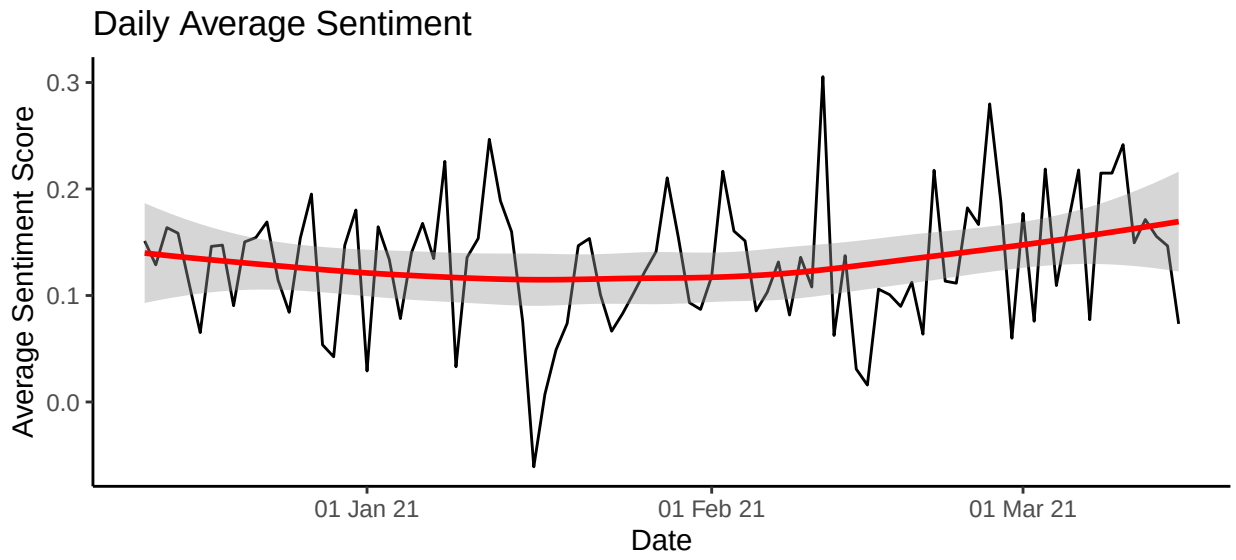


From the bar plot, it was pretty clear that the count of Positive sentiment score (2918) was the highest and the count of Negative sentiment was the least (1296). The count of Neutral sentiment was 2604. This

supports the first hypothesis that the dominant sentiment about the vaccine is Positive which means that people are accepting of the Pfizer-BioNTech COVID-19 vaccine.

Hypothesis 2: The daily average sentiment follows an upward trend

To test this hypothesis, daily average sentiment score was plotted against the date.



As can be seen, from Dec 2020 to Jan 1 2021, daily average sentiment score was going down. It can be observed that the average sentiment had remained positive throughout. We know that the daily vaccinations from all over the world just started in Dec 2020, starting with essential personnel in healthcare being vaccinated. We observed that the average sentiment score from Jan to Feb 2021, remained positive and almost constant. There was a rapid growth in the daily vaccinations during the month of Jan and Feb 2021, but we did not see as much positive change in the sentiment score as that of the increase in the daily number of vaccinations. We saw an increase in the sentiment score starting from Feb 2021 and also in the daily vaccinations doses given. Overall, the strength of average daily sentiment seemed to follow an upward trend since Feb 2021 which supports our hypothesis.

The first two hypotheses tell us that people are supportive of Pfizer-BioNTech COVID-19 vaccine.

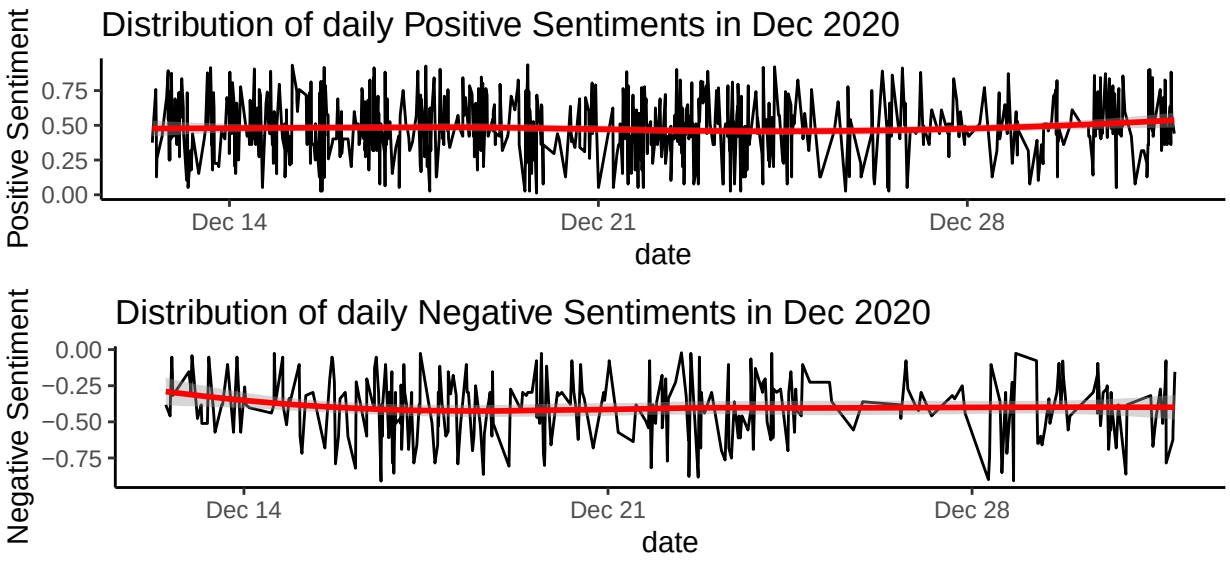
Hypothesis 3: The sentiment series is stationary and can be used to forecast average sentiment score in the future.

To test this hypothesis, the steps given below were followed.

Step 1: We made three partitions of the dataframe as follows:

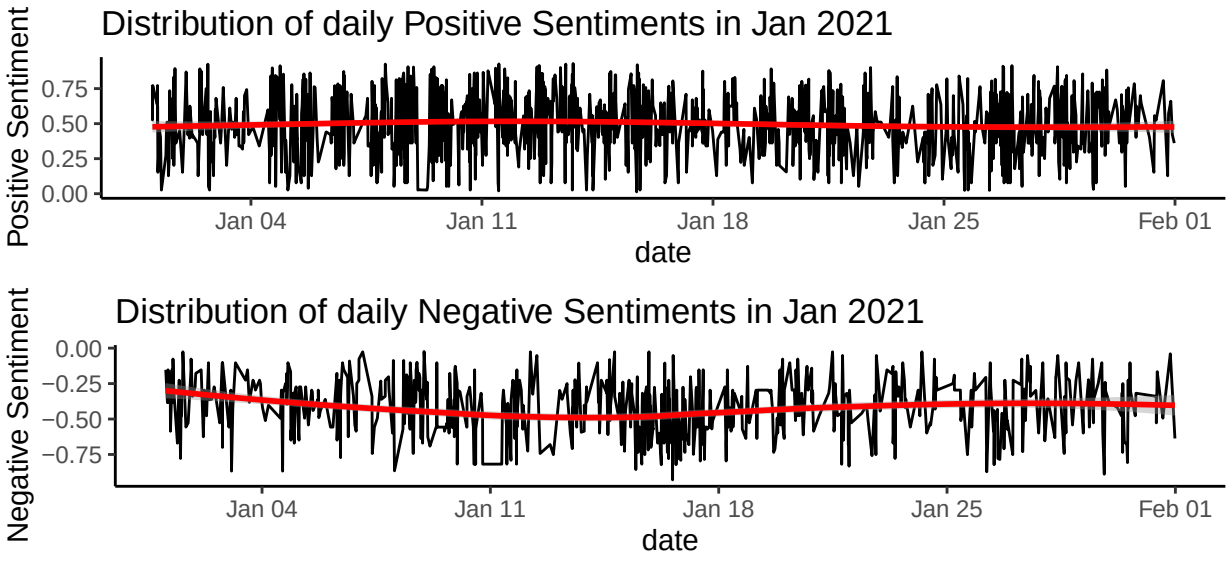
Partition 1: Tweets from 2020-12-12 to 2020-12-31. Partition 2: Tweets from 2021-1-1 to 2021-1-31. Partition 3: Tweets from 2021-2-1 to 2021-3-15. Since we had only 15 days data for Mar 2021, it's data was added to partition3 along with Feb 2021 data.

Step 2: Plotted Positive and Negative Sentiment Score time series for Partition 1 i.e. for Dec 2020



We observed an increase in the negative sentiment strength from Dec 14 to Dec 21, after which it remained almost constant. In the preliminary analysis, we saw that the number of people vaccinated was very low up until the end of Dec 2020 all over the world. The vaccinations started at the end of Dec in many countries and vaccines were initially offered only to essential personnel instead of the general public. So this explains the overall sentiment in Dec 2020.

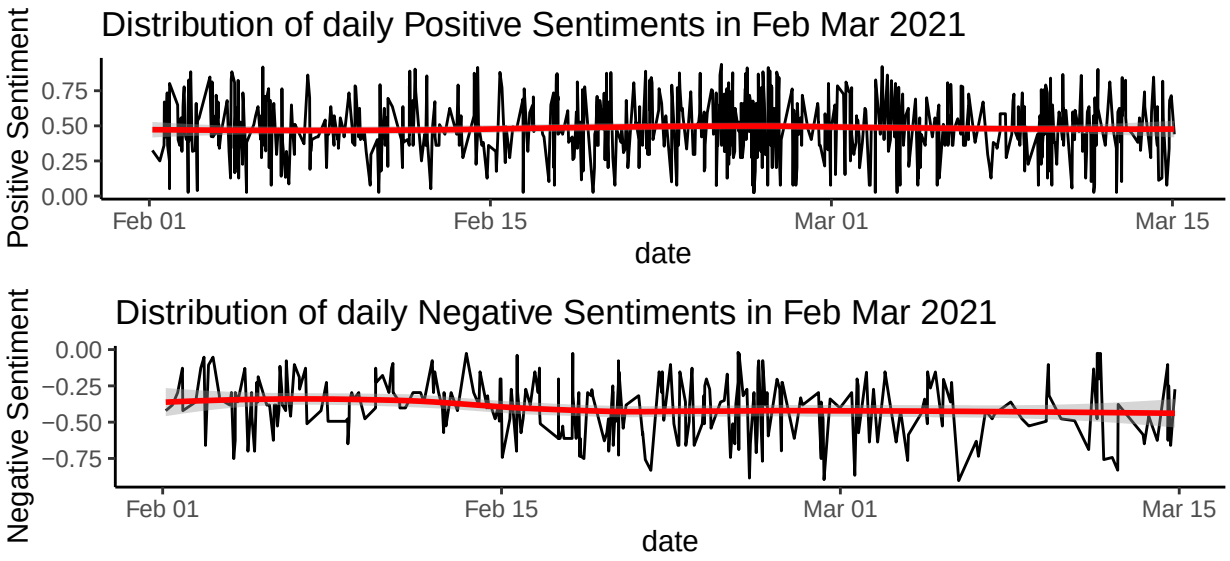
Step 3: Plotted Positive and Negative Sentiment Score time series for Partition 2 i.e. for Jan 2021



From 1st to 31st Jan 2021, the positive sentiment was almost constant. The negative sentiment increased drastically up until Jan 14 i.e. the negative sentiment was more negative and decreased a little after Jan 18. It remained almost constant after Jan 20 till the end of month. According to news, it was around this time

that Scientists feared an "escape mutant" identified in the coronavirus variant first spotted in South Africa which might decrease vaccine efficacy. With the change in the administration at the White house people seem more hopeful which explains the slight decrease in the Negative sentiment strength after Jan 18.

Step 4: Plotted Positive and Negative Sentiment Score time series for Partition 3 i.e. for Feb-Mar 2021



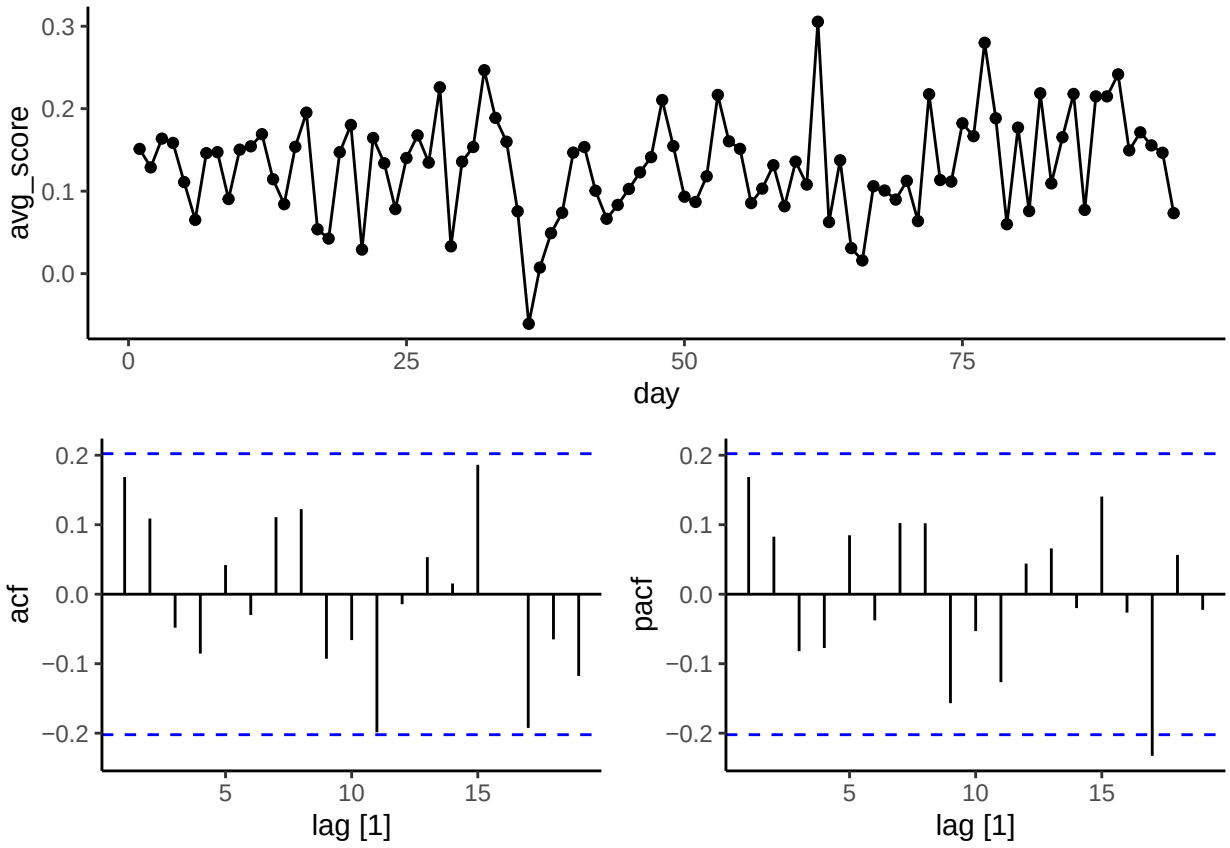
From 1st Feb till 15th Mar, the positive sentiment appeared to be constant. Negative sentiment strength increased till Feb 15 and was almost constant after that. As vaccinations in the country increased, people were concerned about side effects. This provides an explanation for the increase in the negative sentiment strength during this period. With the daily infection rates declined, we can see that the negative sentiment strength stays the same towards the end.

It is important to note that overall average sentiment still remained "Positive" through all three partitions i.e from Dec 2020 till 15 Mar 2021. These time series were analyzed to check how sentiments changed over time and what were some important events that affected these sentiment strengths.

Step 5: ACF Plot & ARIMA Forecasting

Autocorrelation measures the extent of a linear relation between lagged values of time series. The autocorrelation coefficients make up the autocorrelation function or ACF. The plot known as a correlogram is used to plot the ACF to see how the correlations change with lag k. The time series is stationary if it has no autocorrelations. For a stationary series, we expect 95% of the spikes in the ACF plot to lie within the confidence bounds (the blue dashed lines in the plot).

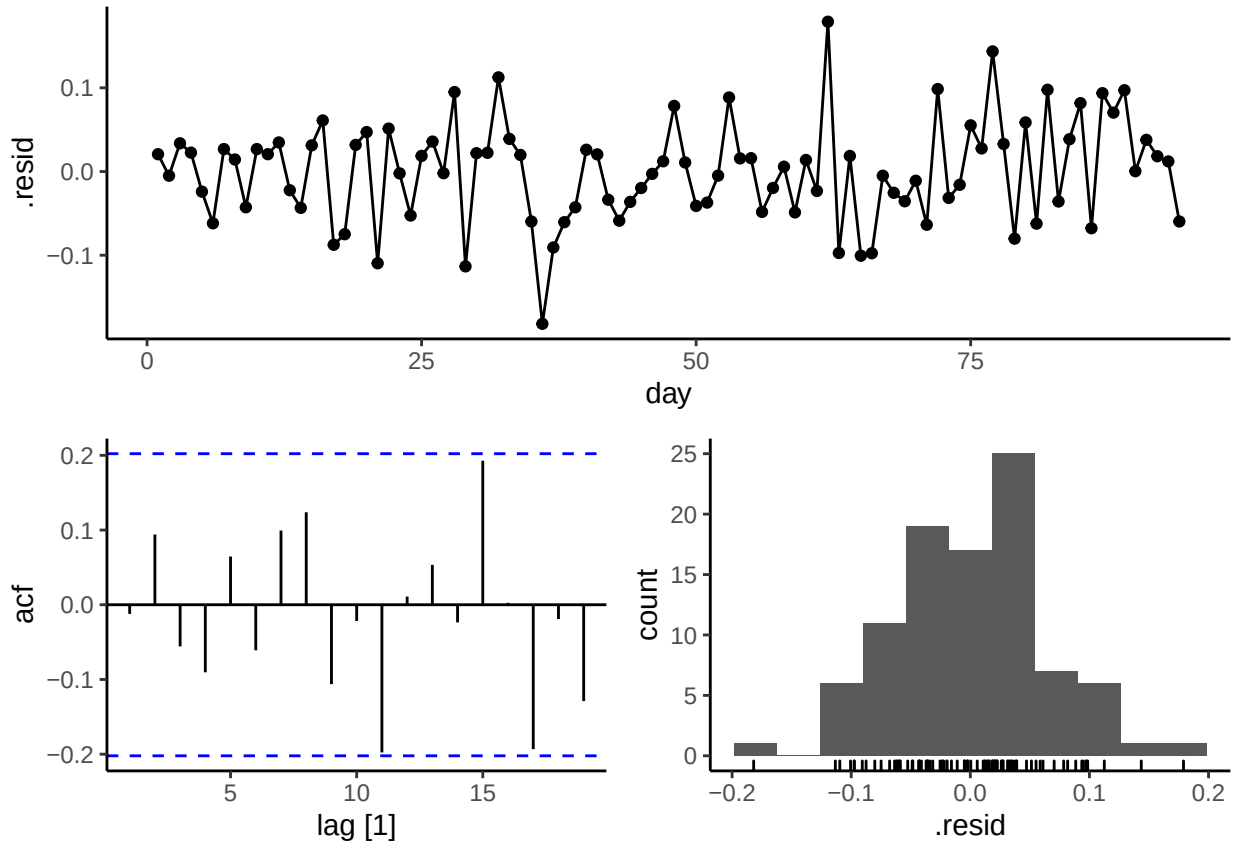
We first created a tsibble object and made calendar adjustments. We then analyzed the ACF,PACF plots for the time series with average sentiment score.



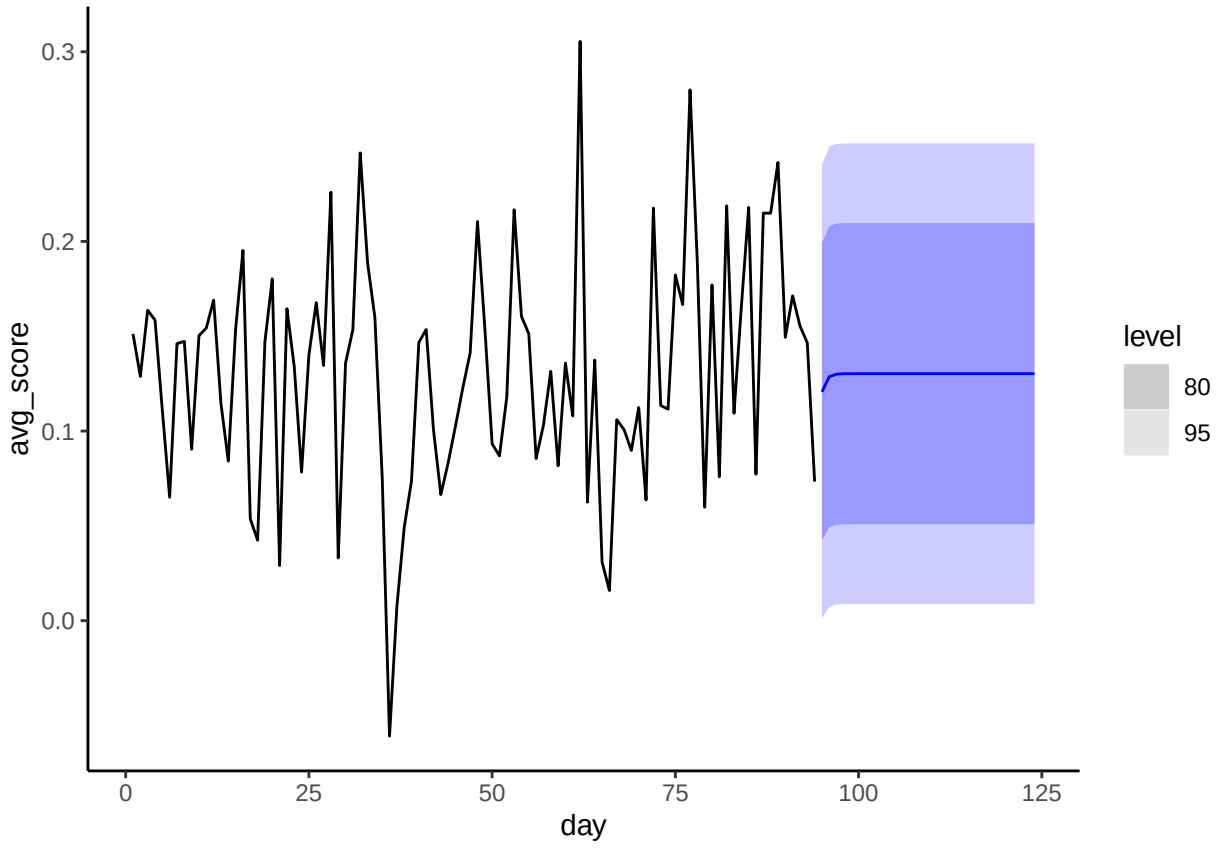
The ACF plot shows that all the lags were within the confidence bounds which meant the time series is stationary. The autocorrelation is close to zero. The PACF plot has one spike at lag 17 extending the confidence bounds but that is still within the 5% tolerance limit. So this shows no partial autocorrelation as well.

We then fitted the ARIMA model to generate forecast for next 30 days.

```
## Series: avg_score
## Model: ARIMA(1,0,0) w/ mean
##
## Coefficients:
##          ar1
##       0.1686
## s.e.  0.1016
##       constant
##         0.1083
## s.e.    0.0062
##
## sigma^2 estimated as 0.003729:  log likelihood=130.42
## AIC=-254.84   AICc=-254.57   BIC=-247.21
```




```
## # A tibble: 1 x 3
##   .model     lb_stat
##   <chr>        <dbl>
## 1 ARIMA(avg~    6.69
## # ... with 1 more
## #   variable:
## #   lb_pvalue <dbl>
```



The ARIMA function chose the model (1,0,0) w/ mean. The ACF plot here tells us that the autocorrelation between any of the residual lags is not significant. These are between the 95% confidence bounds. This means that the residual is white noise or stationary series. The forecast for next 30 days, shows that the strength of average sentiment score will remain constant. This means that the average sentiment score which is Positive in the present case, will remain so in the future. In simpler words, the willingness to get vaccinated will remain constant in the near future.

2. What motivates people to get vaccinated?

Intuitively, we would expect people to get motivated to get the vaccine as vaccinations progress across the world. Also, it would be interesting to see if higher number of daily tweets bolster positive sentiment about the vaccine among people. To check if these intuitions are correct next two hypotheses were tested.

Hypothesis 4: There is a strong correlation between the average sentiment score and the percentage of people vaccinated.

To test this hypothesis, Pearson correlation test was used. The null hypothesis for pearson correlation test was : The true correlation between average sentiment score and average percentage of people vaccinated daily, is equal to 0.

```
## [1] 0.1715598
```

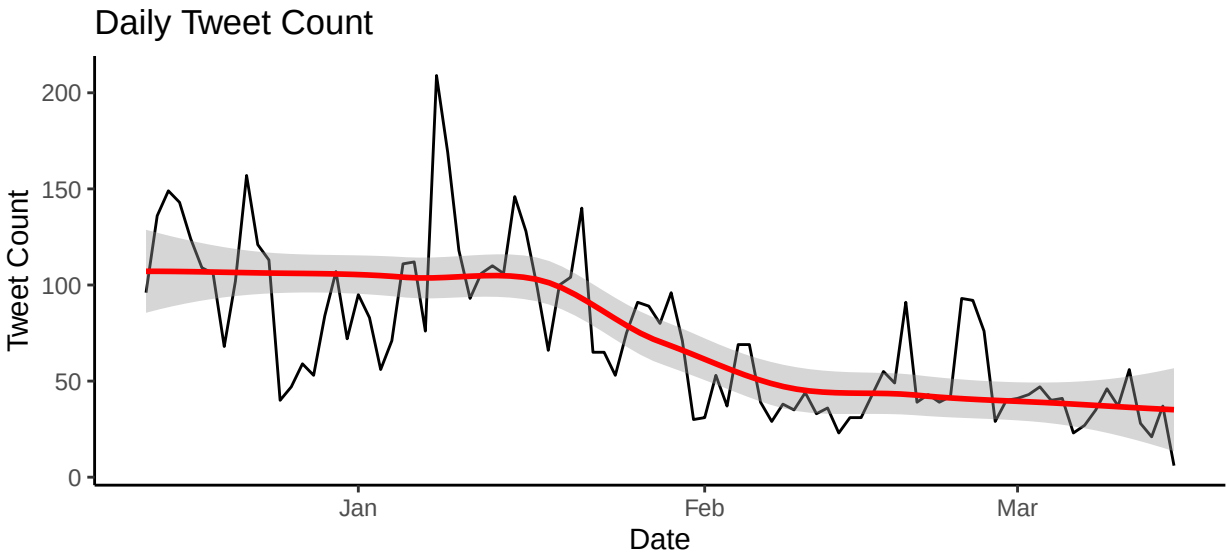
```
##
##  Pearson's
##  product-moment
```

```
##  correlation
##
## data:  total$avg_score and total$avg_percent_people
## t = 1.6612, df =
## 91, p-value =
## 0.1001
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03331265  0.36259605
## sample estimates:
##       cor
## 0.1715598
```

From the output, it can be seen that the p-value 0.10 is greater than the significance level 0.05.Thus, we do not reject the null hypothesis. We conclude that our hypothesis is wrong. There is no strong correlation between the average sentiment score and the percentage of people vaccinated. This is also supported by the low correlation coefficient (0.17).

Hypothesis 5: There is a strong correlation between the average sentiment score and daily number of tweets.

To test this hypothesis,first the daily tweet count trend was analyzed and then Pearson correlation test was used.



We saw spikes in the daily tweets at the following dates which can be linked to important events that happened on that day. 8 Jan 2021 - Commission proposes to purchase upto 300 million additional doses of BioNTech-Pfizer vaccine.

20 Jan 2021 - The presidency of Joe Biden began.

29 Jan 2021 - vaccine found to be effective against variant discovered in U.K.

19 Feb 2021 - Israeli study finds Pfizer vaccine 85% effective after first shot.

Pearson Correlation Test: The null hypothesis for pearson correlation test is : The true correlation between average sentiment score and daily no. of tweets, is equal to 0.

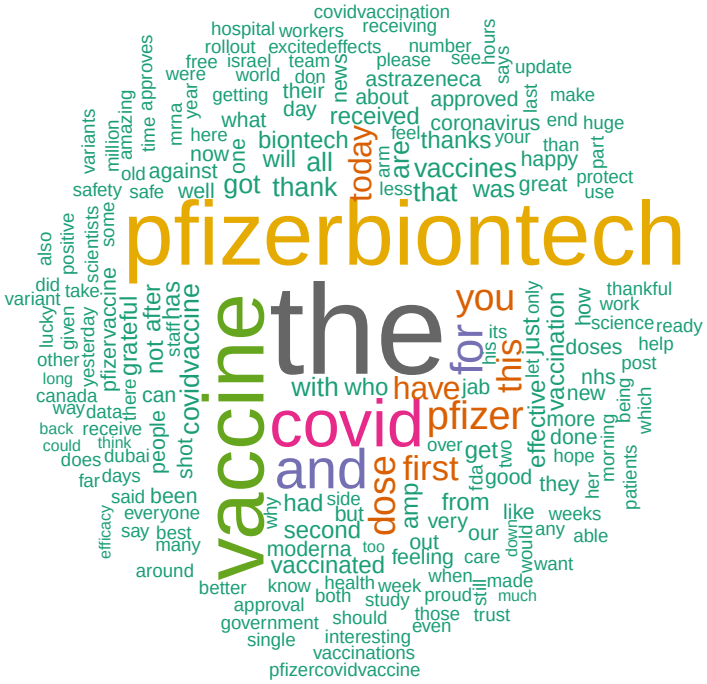
```
## [1] 0.002633575
```

```
##
##  Pearson's
##  product-moment
##  correlation
##
## data:  total$avg_score and total$n
## t = 0.025123, df
## = 91, p-value =
## 0.98
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2011826  0.2062312
## sample estimates:
##         cor
## 0.002633575
```

The p-value 0.98 which is greater than the significance level 0.05. Thus we do not reject the null hypothesis. We conclude that our hypothesis is wrong. There is no strong correlation between the average sentiment score and daily number of tweets. This is also supported by the low correlation coefficient we got (0.002).

We understood that number of people vaccinated and daily number of tweets have no impact on the average sentiment score from hypotheses 4 & 5. We then analyzed what words are frequently used by people whose tweets had a positive sentiment score.

Analyzing Tweets having a Positive Sentiment

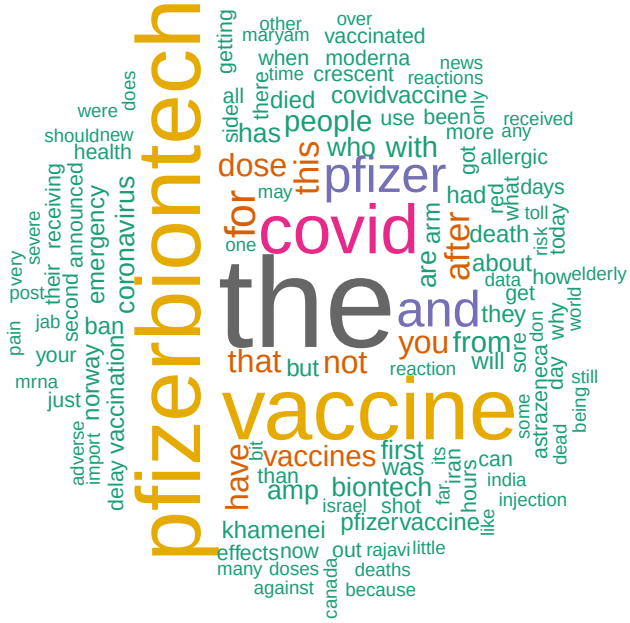


Let us ignore the most frequent words such as pfizerbiontech,vaccine, covid etc. We see some interesting words like scientists, approvals, study,effective, better, positive etc along with some positive words like grateful, great and positive. In this case, we can draw an inference that people having positive sentiment i.e. the people supportive of Pfizer-BioNTech COVID-19 vaccine were influenced by the information about the vaccine studies conducted, approvals given and scientific data available. They also seem to be the people who have received at least first dose of the vaccine.

3. What are the top reasons for the public hesitancy about getting vaccinated?

We are interested in understanding why people are hesitant or not bothered about the Pfizer-BioNTech COVID-19 vaccine so that we can propose some strategies to change their opinion about the vaccine.

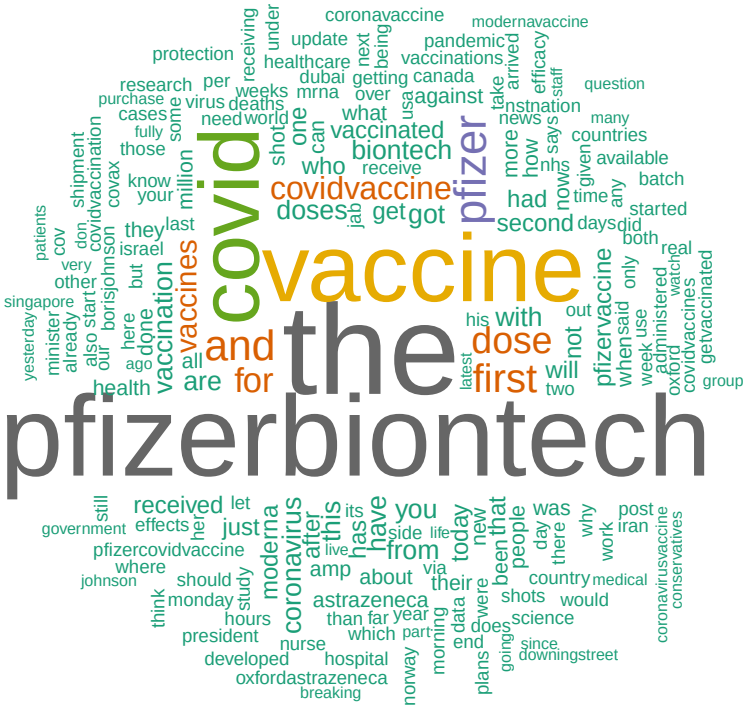
Analyzing Tweets having a Negative Sentiment



When we look at the wordcloud of negative sentiment tweets, we see alarming words such as allergic, emergency, severe, side effects, deaths. This gives us an idea that people having negative sentiment about the vaccine are worried about the vaccine side effects, whether it causes death etc. We also see some words like ban and injection. This can be used to let people know about the approvals and explain the severity of side effects compared to actual covid-19 symptoms.

G. Future Opportunities :

Analyzing Tweets having a Neutral Sentiment



When we look into the wordcloud of neutral sentiment tweets, we see other vaccine names like moderna, covax , astrazeneca.This information can be used by Pfizer to promote their vaccine. Most of the other words are similar to those of negative sentiment tweets.

G. Future Opportunities :

While this analysis may seem thorough, there are a few opportunities to improve this analysis further. These are listed as follows:

- As can be seen from the wordcloud of people having Neutral sentiment, Pfizer-BioNTech as an individual brand has not much impact. Pfizer-BioNTech can use this analysis to improve their public reach and emphasize how their vaccine is better as compared to other vaccine manufacturers.
- This analysis can be extended to other vaccines used in USA like Moderna.
- Vaccine manufacturer's can approach people for next clinical trials if their overall sentiment about the vaccine has been positive.

H. Conclusion :

Now that we have reached the end, it is important to summarize and conclude the hypotheses testing and sentiment analysis. Using the Pfizer and BioNTech Vaccine Tweets and Vaccination progress datasets, we conclude:

1. The willingness to get vaccinated has been increasing, however, in future the willingness seems to remain constant.
2. The percentage of people vaccinated and daily number of tweets do not motivate people to get vaccinated. What motivates them is the scientific data, information about the approvals given for the vaccine and efficacy of the vaccine. We also see that people who got vaccinated are more confident about the vaccine. These people can be approached to advocate about the vaccine through social media.
3. Some of the reasons for the public hesitancy towards Pfizer-BioNtech and vaccines in general, are concerns like safety, severe side effects and allergies. Thus, the public health authorities should design and plan their campaigns in such away that they emphasize more on the concerns people have. They should provide more information and transparency on the efficacy and approval process through social media so as to reach maximum public.

I. References and Sources :

1. Pfizer and BioNTech Vaccine Tweets - This dataset has tweets about Pfizer-BionTech Vaccines. https://www.kaggle.com/gpreda/pfizer-vaccine-tweets
2. COVID-19 World Vaccination Progress - This dataset has data about Daily and Total Vaccination for COVID-19 in the World https://www.kaggle.com/gpreda/covid-world-vaccination-progress/code?datasetId=1093816&language=R
3. https://en.wikipedia.org/wiki/COVID-19_recession#:~:text=The%20COVID%2D19%20pandemic%20is,30%20Januar
4. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/how-they-work.html
5. https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-confirm-high-efficacy-and-no-serious
6. https://www.theguardian.com/world/2021/feb/25/pfizer-covid-vaccine-94-effective-study-of-12m-people-finds