010 Element ASSIGNMENT

APPLIED DATA ANALYSIS AND RESEARCH METHODOLOGY

MOD007893

Adrien Montaigne
2248806

Contents

List of figures

I. Abstract

The vaccination makes it possible to fight against the COVID-19. The more people are vaccinated, the more effective the vaccination is. In the current study, we examined how different vaccines were perceived by the public on social media, notably Twitter. To understand how people, feel about vaccination, we employed a variety of tools, including VADER and TextBlob, as well as two separate datasets from Kaggle. The project's goal was to measure and quantify the sentiments expressed in tweets and contrast those measurements with various attributes (the popularity, the date, the source, ...). Our analyses show that there is little correlation between the public's opinion on the vaccine and the difference in feature, except for the hashtags. The population did not hold the same opinion for each vaccine. Oxford-AstraZeneca and Sinopharm were less well received than Pfizer.

II. Introduction

1. COVID

Beginning in 2020 COVID-19 appeared in China, and it quickly spread throughout the entire planet. Many countries were forced to close their borders and applied a lockdown. The lockdown was the first measure to fight against COVID. Financially and socially heavy, most countries tried different measure such as masks or vaccination. The vaccines though are effective only if the majority of people decide to be vaccinated.

2. Vaccination

The pharmaceutical firms in different countries developed different vaccines, below is a list of the most common vaccines now in the world:

- Pfizer/BioNTech (United States)
- Moderna (United Kingdom)
- Sinopharm (China)
- Oxford-Astrazeneca (United Kingdom)
- Sputnik (Russia)
- Covaxin (India)
- Sinovac (China)

The effectiveness is different for each vaccine. The best efficacy rate is shared by Moderna and Pfizer. Each vaccine also has a different secondary effect.

3. Twitter

Twitter is a social network that people use all throughout the world, excluding China. Twitter allows the user to express their opinion through a tweet, a message that contains less than 160 characters. Due to the popularity of Twitter, the prevalence of fake news, and unending rumours on the platform, we decided to analyse those tweets.

4. NTL (Natural language Technology)

Natural Language Technology refers to the interaction between a natural language human and a computer. In the social network, NLT is widely prevalent. It seeks to comprehend how people think, feel, and behave. It is widely used in marketing intelligence, sentiment analysis, customer service, and advertising. We used VADER and Textblob for sentiment analysis in this project. Below is a list of the various tasks that NLT can complete:

- Speech–to-text
- Automated translation tools (google translation or Deepl)
- Chatbots
- Social media monitoring

5. The objectives

We will analyse the sentiment of the population on the different vaccines on various parameters: hashtags, popularity, whether the users are verified to identify top user's characteristics and tweet's engagement in each vaccine opinion.

III. Data pipeline attributes

1. Dataset

5 V's (Veracity, Volume, Variety, Veracity, Value) are the five main characteristics for a dataset. In this report, we used two structural datasets: vaccination_tweet.csv, a dataset on Pfizer vaccine tweets, and vaccination_all_tweets.csv, a dataset on all vaccine tweets. Both vaccination tweet.csv and vaccination all tweets.csv were created by Gabriel Preda, a data scientist working for Endava. The link is in reference. vaccination_tweet.csv and vaccination_all_tweets.csv have exactly 16 columns with the same feature: ID, user name, user location, user description, user created, user's followers, user friends, user favourites, user verified, date, text, hashtags, source, retweets, favourites and is retweets. The datasets record the tweets between December 2021 and April 2022. The time frame coincides with the start of the vaccination. We selected those datasets because we wanted to compare the views on various vaccinations. Thus, if one vaccine has a better reputation than another, we could interpret that. The volume of the first dataset (Pfizer vaccine tweets) is 2 MB and the volume of the second of the dataset (All vaccine tweets) is 31 MB. The first dataset's shape is (8082, 16), and the second dataset's shape is (228207, 16). second is larger than the first one. The most crucial column is text since we will change the text into sentiment and contrast it with other features. To obtain reliable data, we have selected two large datasets. Twitter has a reputation for being a social media platform where users can openly express their opinions. LinkedIn and Facebook, in comparison, are social networks where individuals express their thoughts less.

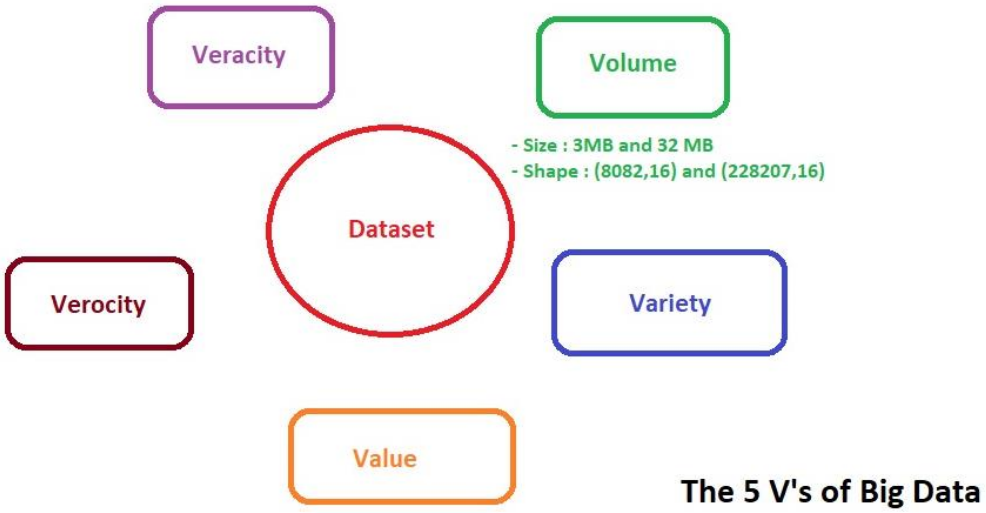

***Figure 1**: The 5 V's of Big data*

2. Data exploration

Data exploration permits us to know more information about the dataset such as its shape, the type of columns, the number of rows, the average, or the medium of all columns. Data exploration permits to identify the missing values or the duplicate values.

3. Data cleaning

For this project, we used pandas, numpy, nltk for VADER, sklearn, re, matplolib, tabulate, seaborn, and etc. We removed duplicate variables and dropped the columns that were not important for the project. Then, we encoded the columns as numbers: hashtags, source, is_retweet, because the variables are unreadable for the analysis. For example, replacing True and False by 0 or 1.

IV. Design and development (methodology)

1. Data Pre-processing

We filtered all emojis, special characters, links, hashtags ... (remove all that does not correspond to a word). We convert every character to lower case. That might be helpful if we wish to analyse public sentiment over time. Then, you have to extract the dates: transform the column containing dates of format DD/MM/YYYY H:MM: SS AM/PM into several columns: a column for the month, the year and the day. This will allow us to compare feelings and events related to the vaccination. The second step is to create four columns: Positive score, Negative score, Neutral score and Category to measure the sentiment for each tweet with VADER. The third step is to create three columns: Polarity, Subjectivity and Category to measure the sentiment for each tweet with TextBlob.

2. VADER (Valence Aware Dictionary for sentiment reasoning)

VADER (Valence Aware Dictionary for sentiment reasoning) permit to distinguish the sentiment (neutral, positive, negative) and the intensity (between -1 and 1) of the sentiment from a text. VADER combine a lexical dictionary and 5 heuristics.

VADER was created by a few researchers in Poland, in 1983. VADER is based upon the idea of wisdom of the crowd. The researchers ask what the public feels about lexical feature in responding by a few words like “ok”, “good”, “great” for positive sentiment or like “horrible”, “sucks” for negative sentiment. Each word corresponds to valence score. For example, the positive score for ok is 0.9. For each lexical feature, the researchers sum the valence score and normalized the variable between -1 and 1 to provide a single unidimensional measure of sentiment. The 5 Heuristics consists to intensify in function of the punctuation, the capitalization, the shift in polarity due to “but”, the previous sentences or words or the degree of the modifiers like for example effing cute or sort of cute. VADER's strategy is well suited for platforms such as Twitter. VADER doesn’t require training data and it’s a fast method. VADER needs a large dataset like the one we have because it can understand trends well. VADER's lack of sarcasm recognition makes it susceptible to misreading messages and misunderstandings. The neutral emotion, the positive feeling, and the negative feeling will each have their own column in this section.

3. TextBlob

TextBlob is a library python for sentiment analysis like VADER. TextBlob permit to measure the polarity between -1 and 1 (-1 means that the text is extremely negative. In contrast, 1 means that the text is extremely positive) and the subjectivity between 0 and 1 (0 for objectivity and 1 for subjectivity). The critical difference between VADER and TextBlob is that VADER is much more suited for social media. VADER puts a lot of effort into identifying the sentiments of content that typically appear on social media, such as emojis, repetitive words and punctuation.

4. Flow Chart

The diagram below represents the steps of my project.

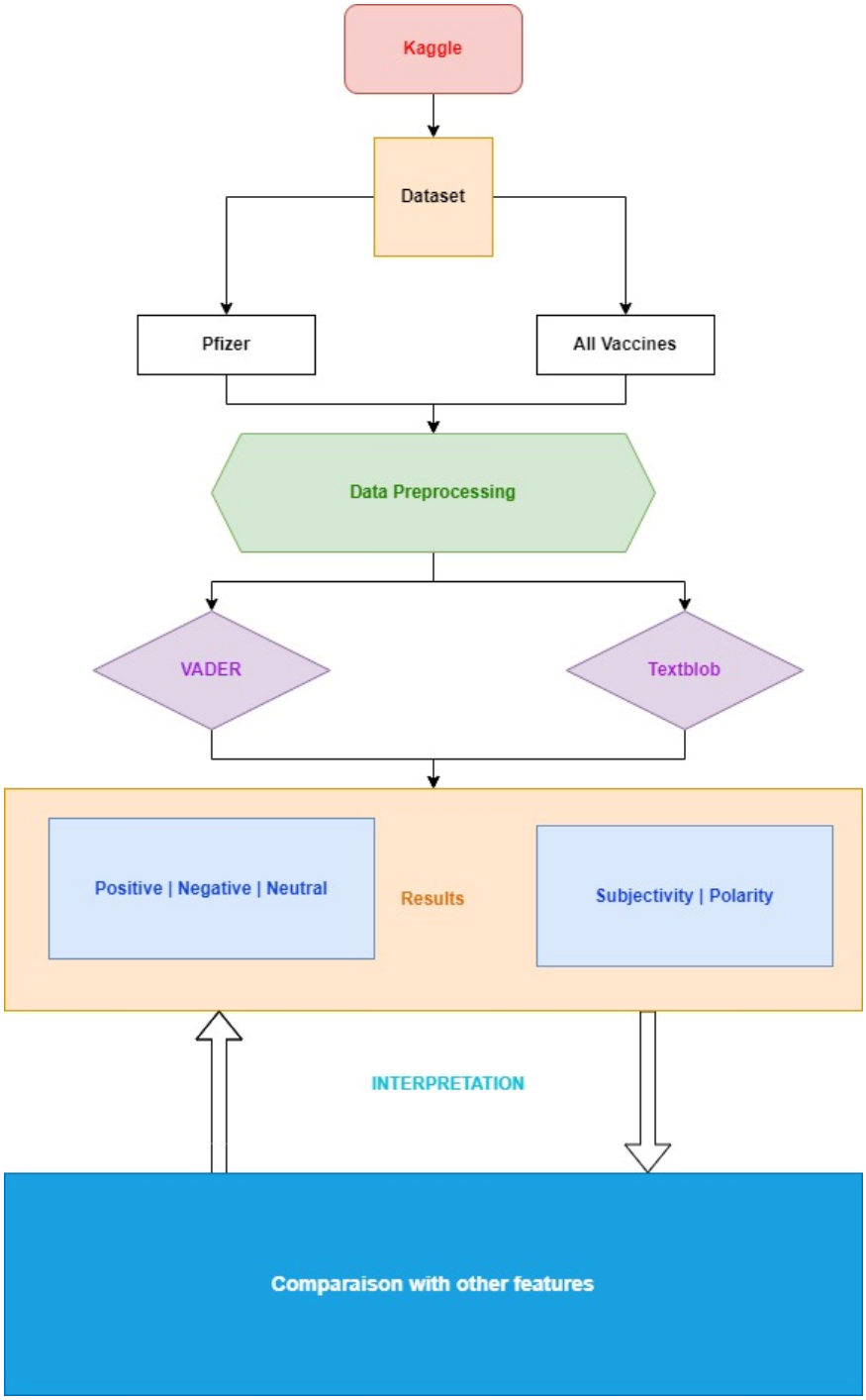

Figure 2: *Diagram on the step of my project*

V. Data visualization and storytelling

1. Proportion of Neutral, Positive and Negative tweets

From tweets cleaned, we will create four columns with the help of VADER: negative score, positive score, neutral score, and the category of the sentiment. We could observe on the diagram that the most of tweet are Neutral or Positive.

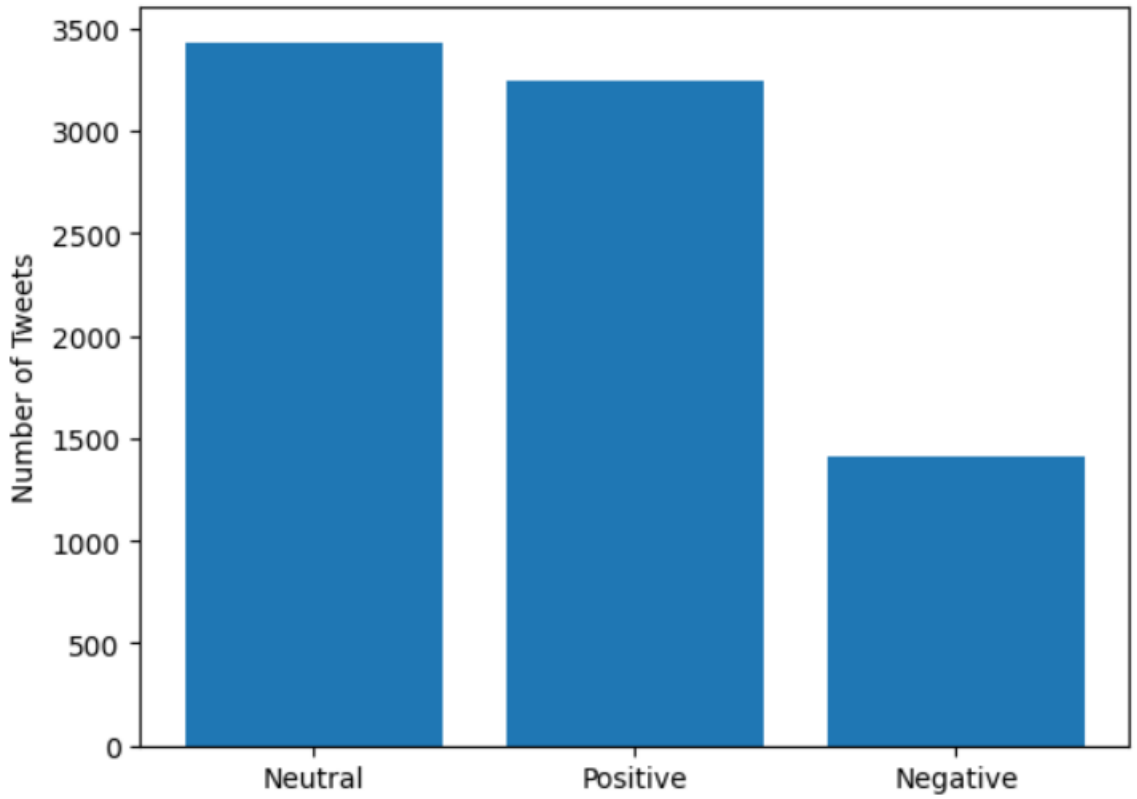

Figure 3: *Diagram on the Number of Positive, Neutral, and Negative tweets*

For the first dataset, the average score valence for neutral emotion is 0.82. The score average for positive sentiment is 0.12 and for negative sentiment is 0.05.

2. Sentiment analysis according to the time

However, the general perception of vaccinations varies on the circumstances at the time, the days, or the months. This histogram shows that people generally have lower opinions in January than they do in other months. The reason is that the beginning of vaccination campaigns worldwide coincides with January. At first, people may feel anxious about the new vaccine.

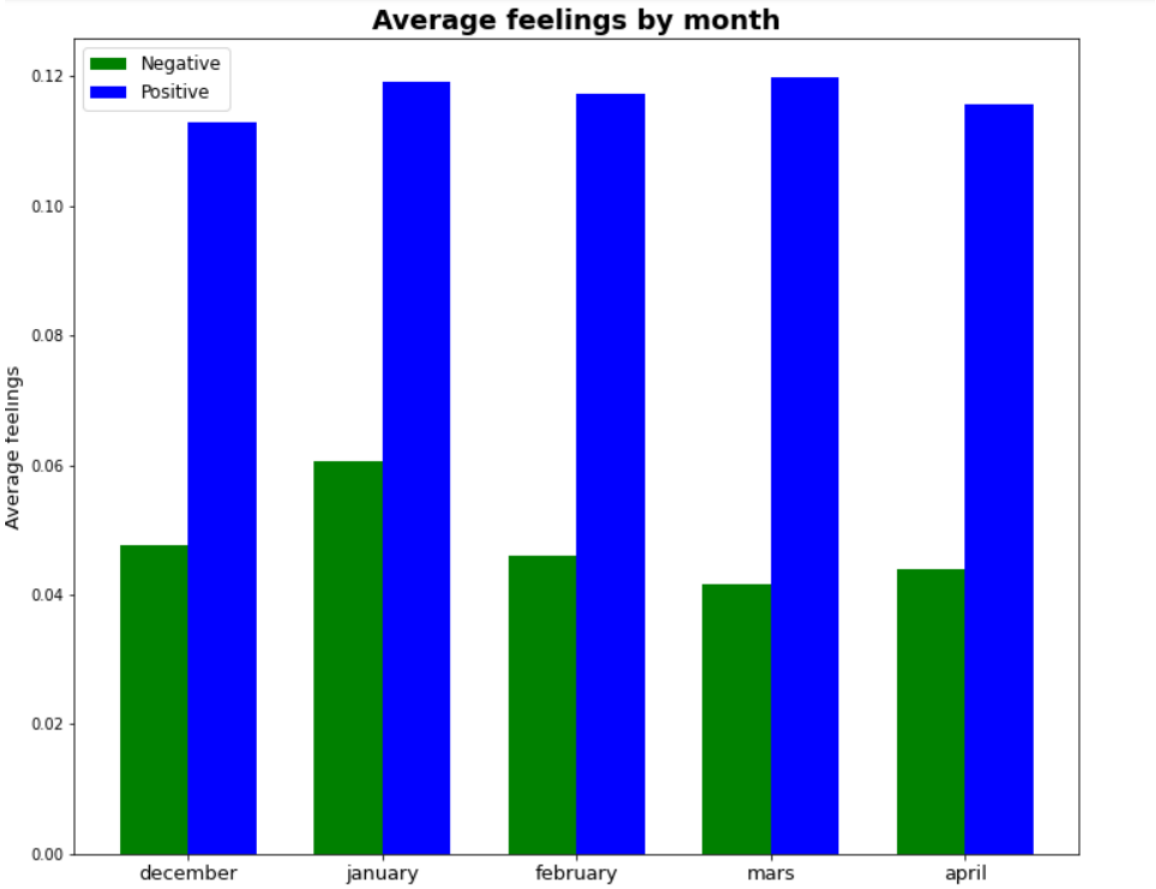

***Figure 4**: Graphics on the public opinion in function the month*

The United Kingdom is the first western country to accept the vaccination, the 2 December 2021. France starts the vaccination campaign, the 28 December December 2021. The first vaccination is administered in India on January 16. In January 2022, the vaccine campaign will be launched in many nations around the world. We suppose that the negative tweets are correlated with the beginning of the vaccination campaign.

3. Characteristics of user

We now wondered who propagated these tweets with a negative sentiment. Firstly, the majority of people are unpopular on Twitter and the most of accounts are not verified. We could observe on the donut chart that 90% of the users have an unverified account (represented by the dark blue). More than 50% have less than 500 followers (represented by the light blue on the pie chart. Only 10% have more than 10 000 followers (represented by the dark blue on the pie chart). Figure 5 is a donut chart on the feature: user verified. Figure 6 is a pie chart on the user popularity. Both figures were created by Power BI.

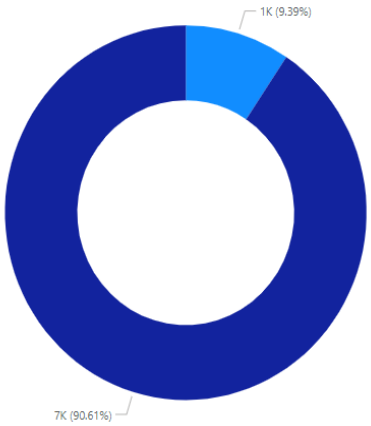

Figure 5: *Donut chart on the feature: user_verified*

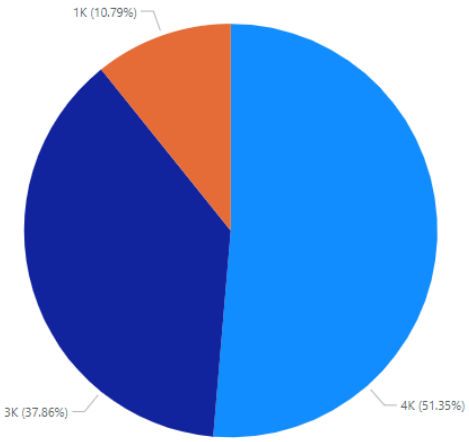

Figure 6: *Pie chart on the user popularity*

Less tweets come from popular accounts than from less popular accounts, on average. However, the effect of a tweet on public opinion is different depending on the popularity. We will wonder what popular user think about the Pfizer or BioNTech. That could give an idea how the fake new or rumours spread.

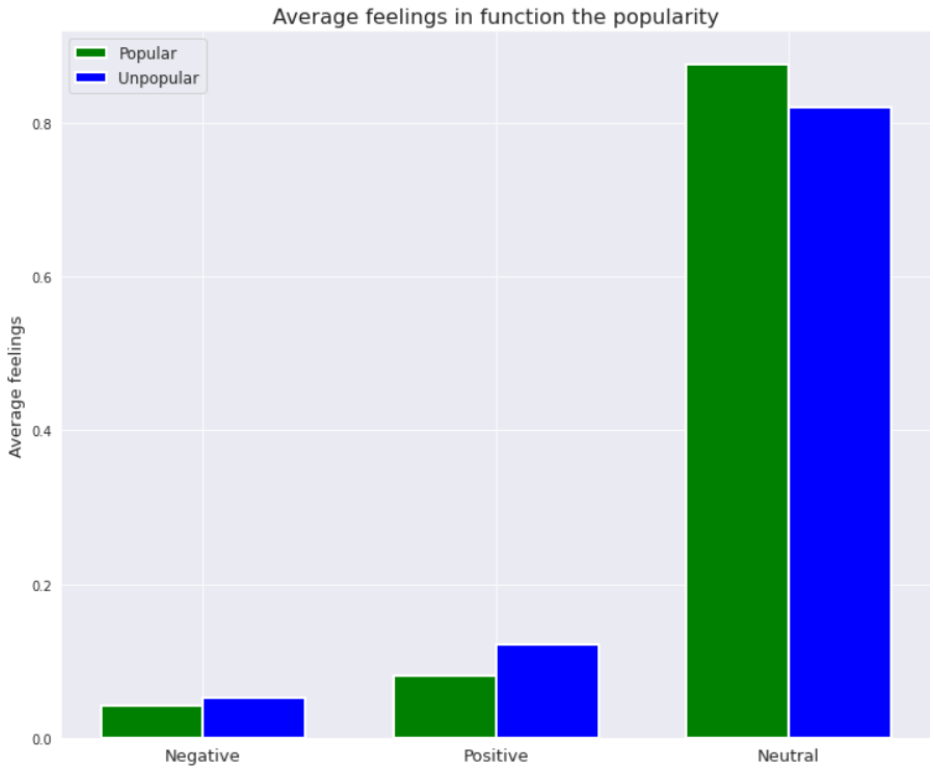

Figure 7: *Histogram on the average feelings in function the popularity*

Popular users tend to tweet fewer positive and negative messages than unpopular users. However more neutral tweets are posted by popular users than by less popular users. The reason

for this may be first-off that well-known users draw attention to their publications and secondly, unpopular user has less limited by the followers.

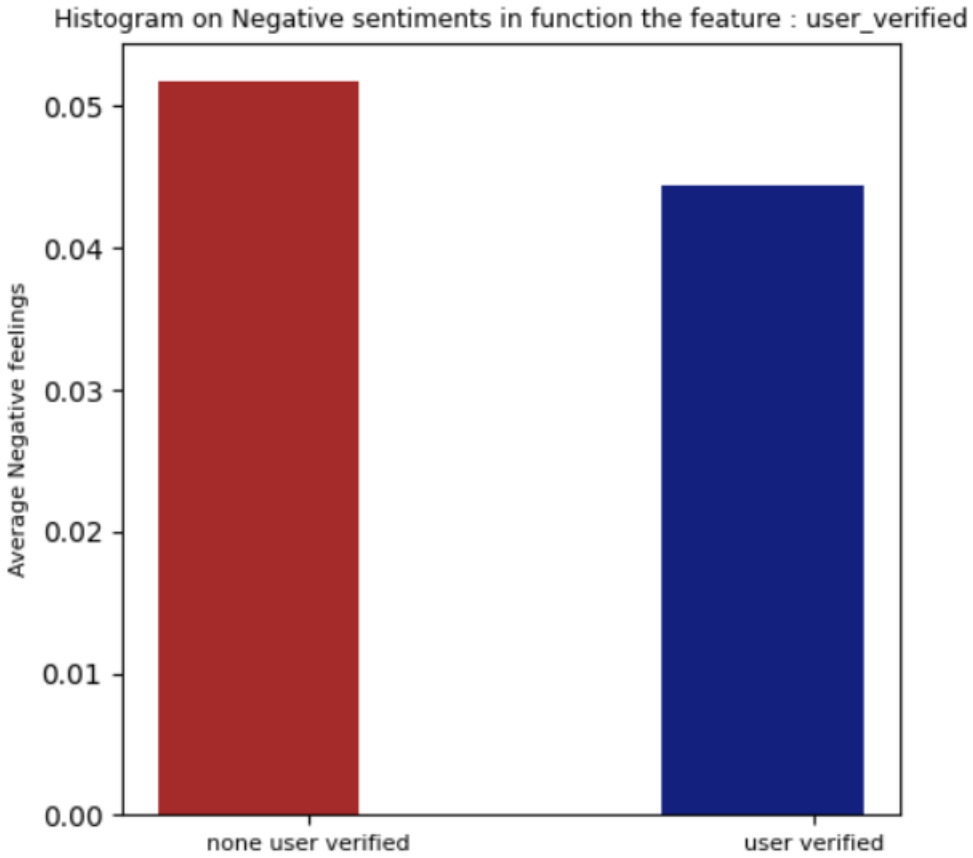

Figure 8: *Histogram on Negative sentiments in function the feature: user_verified*

The histogram shows the valence score for negative sentiments in function if the user has verified their account. User verified tends to have a good opinion on Pfizer 14.259572 more than none user verified. We must therefore pay attention to the interpretation because it's possible that Twitter delete certain type of users or tweets.

4. Comparison between the datasets

Afterwards, the comparison between different dataset is necessary to insure the reliability. The second dataset contains tweet from all vaccinations. We will implement the same pre-processing. Figure 8 represents the comparison table for Positive, Negative, Neutral feelings in function of the dataset.

Positive feeling :

Pfizer	All vaccination
0.117156	0.0981105

Negative feeling :

Pfizer	All vaccination
0.0511137	0.0451675

Neutral feeling :

Pfizer	All vaccination
0.826039	0.850257

Figure 9: *Table comparing Positive, Negative, Neutral according to the datasets*

5. Comparison between the different vaccines

It could be noticed that both dataset's outcomes are remarkably comparable. That corresponds with the previous results. We will now examine how the general population feels about the various vaccines. Figure 10 is a histogram representing the Ratio Negative score / Positive score. Sinopharm are less appreciated by the population that Pfizer.

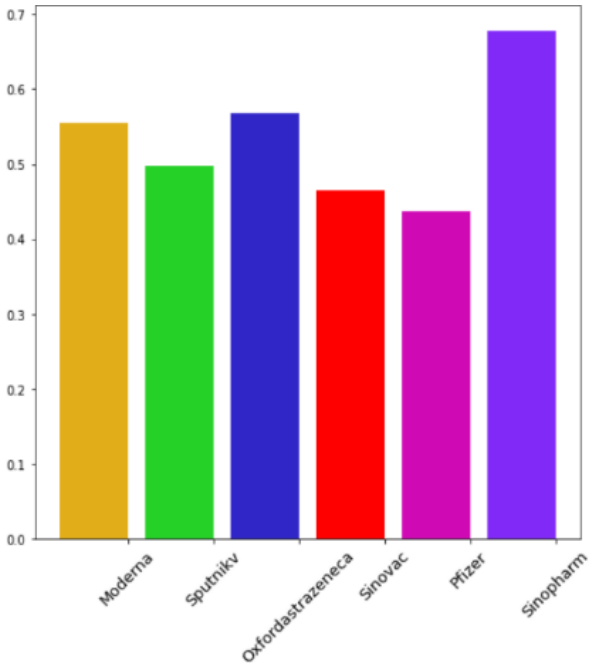

Figure 10: *Histogram on the various vaccines*

VI. Evaluation

1. Word Cloud

Because we can't compare the prediction to reality, evaluating an unsupervised situation is challenging. We used word clouds for this project to check ours results. Figure 10 shows the Word Cloud for common words among positive and negative feelings tweets according to VADER.

Figure 11: *Word Cloud for common words among positive and negative feelings tweets according to VADER*

"Amazing", "proud", "Happy", "Thank" are positive word that included in the positive word cloud. Negative words like "death", "reaction"," ban "and" emergency" are part of the negative word cloud. We could conclude that VADER works correctly.

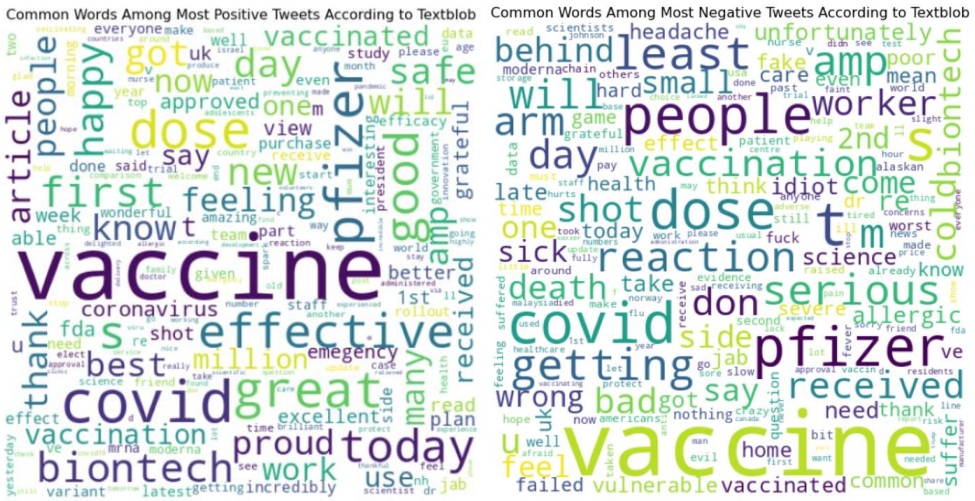

Figure 12: *Word Cloud for common words among positive and negative feelings tweets according to TextBlob*

Figure 11 represents the word cloud for common words among positive and negative feelings tweets according to TextBlob. It observed that the common words between VADER and TextBlob are different. Only few words are identic like “happy”, “good”, today” between the words cloud. after observation we can conclude that TextBlob is working properly.

2. Correlation Matrix and accuracy score

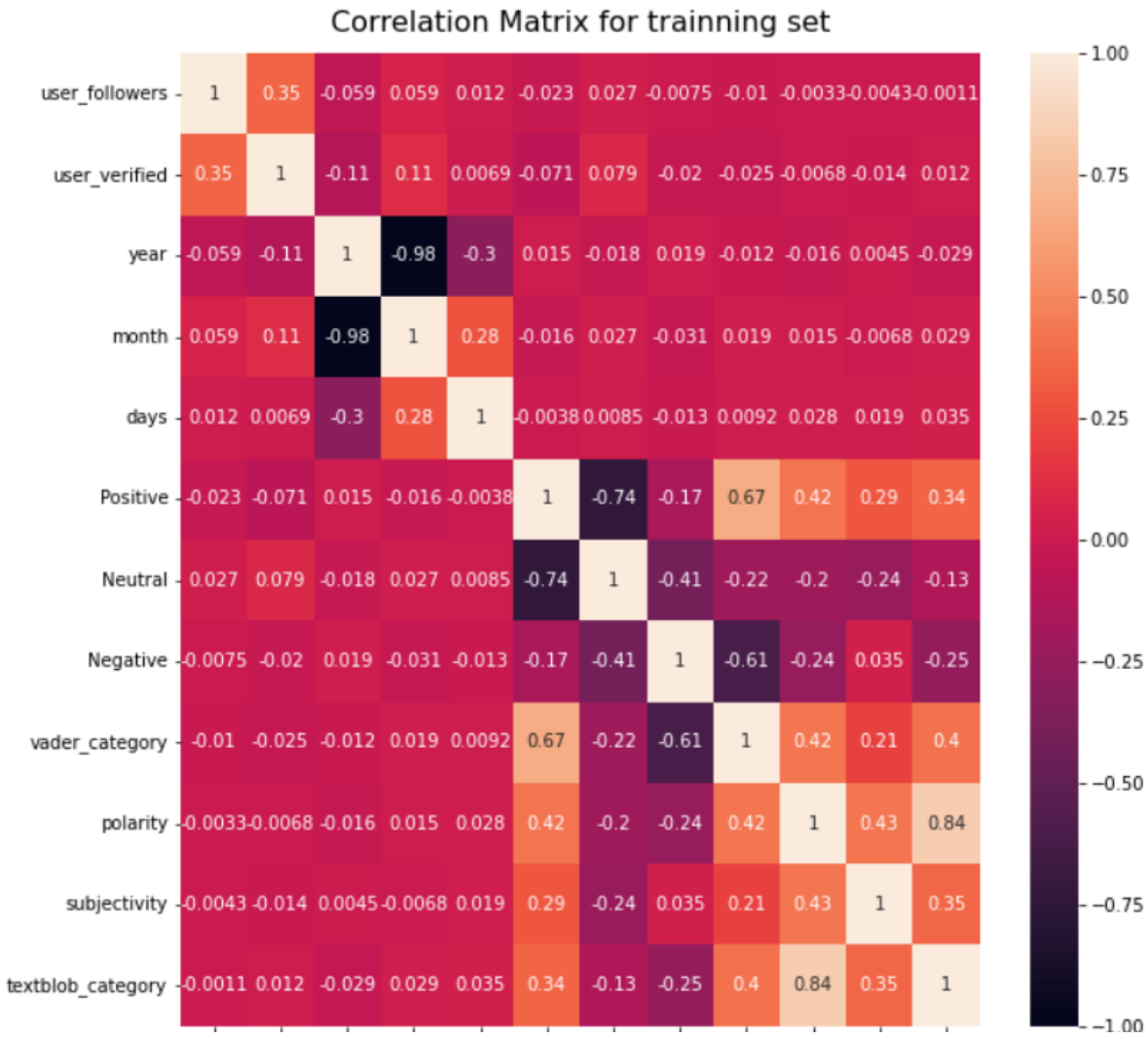

Figure 13: *Correlation Matrix with TextBlob variable*

The comparison of the VADER and TextBlob results permit us to verifies the accuracy of our interpretations. In order to observe the similarity between both models, we use correlation matrix, confusion matrix and accuracy score. The accuracy rating between TextBlob and VADER in this case is 0.58. This indicates that, 58% of the time, TextBlob and VADER predict the same result. "Negative," "Neutral," and "Positive" are the three possible outcomes. Even if sentiment analysis is the topic, 0.58 is a very poor score. We could observe the correlation Matrix on Figure 12. The variable TextBlob category and VADER Category are moderately correlated.

4. Data confidentiality, privacy, and any GDPR issues

Since 2018, Twitter follows the General Data Protection Regulation (GDPR) which is based on existing European Union rules on data protection and privacy. We have to able to delete every time a user or a tweet if necessary.

5. Self-criticism

I encountered several problems before, during, and after the assignment. Before I began the individual study, I was unsure of how to analyse feeling and emotion in a scientific way. I ran into some issues trying to extract date-related data. Because the project was unsupervised, it was challenging to assess the model. The results are difficult to interpret because they are not statistically significant. We are limited by the model and the dataset. Our interpretation depends on a lot of parameters.

VII. Conclusion

Thus, the public opinion on vaccination has evolved over time with different factors and events. The project makes it possible to learn more about the influence of social network on the public opinion and learn more about the tools used in data science. Pfizer was the vaccine most well received by the population. The public's opinion on vaccination had been the lowest point in January, during the beginning of the vaccination campaign. The verified Twitter users tended to get a better opinion on the vaccination in contrast to unverified use. Twitter might censor some tweets. Likewise, with the user's popularity, the known accounts tended to hold a more neutral opinion on the vaccination. Further research should be carried on the topic.

VIII. Reference

Alduaiji, N., Datta, A. and Li, J. (2018). Influence Propagation Model for Clique-Based Community Detection in Social Networks. *IEEE Transactions on Computational Social Systems*, 5(2), pp.563–575. doi:https://doi.org/10.1109/tcss.2018.2831694.

Bang, Y., Ishii, E., Cahyawijaya, S., Ji, Z. and Fung, P. (2021). Model Generalization on COVID-19 Fake News Detection. *arXiv:2101.03841 [cs]*. [online] Available at: https://arxiv.org/abs/2101.03841 [Accessed 1 Apr. 2023].

Blei, D., Ng, A. and Jordan, M. (2023). *Google Scholar*. [online] Google.com. Available at: https://scholar.google.com/scholar_lookup?journal=Journal+of+Machine+Learning+Research&title=Latent+dirichlet+allocation&author=DM+Blei&author=AY+Ng&author=MI+Jordan&volume=3&issue=Jan&publication_year=2003&pages=993-1022& [Accessed 1 Apr. 2023].

Bonnevie, E., Gallegos-Jeffrey, A., Goldbarg, J., Byrd, B. and Smyser, J. (2020). Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *Journal of Communication in Healthcare*, pp.1–8. doi:https://doi.org/10.1080/17538068.2020.1858222.

Du, J., Michalska, S., Subramani, S., Wang, H. and Zhang, Y. (2019). Neural attention with character embeddings for hay fever detection from twitter. *Health Information Science and Systems*, 7(1). doi:https://doi.org/10.1007/s13755-019-0084-2.

Love, B., Himelboim, I., Holton, A. and Stewart, K. (2013). Twitter as a source of vaccination information: Content drivers and what they are saying. *American Journal of Infection Control*, 41(6), pp.568–570. doi:https://doi.org/10.1016/j.ajic.2012.10.016.

Piedrahita-Valdés, H., Piedrahita-Castillo, D., Bermejo-Higuera, J., Guillem-Saiz, P., Bermejo-Higuera, J.R., Guillem-Saiz, J., Sicilia-Montalvo, J.A. and Machío-Regidor, F. (2021). Vaccine Hesitancy on Social Media: Sentiment Analysis from June 2011 to April 2019. *Vaccines*, 9(1), p.28. doi:https://doi.org/10.3390/vaccines9010028.

Steffens, M.S., Dunn, A.G., Wiley, K.E. and Leask, J. (2019). How organisations promoting vaccination respond to misinformation on social media: a qualitative investigation. *BMC Public Health*, 19(1). doi:https://doi.org/10.1186/s12889-019-7659-3.

Zhou, J., Yang, S., Xiao, C. and Chen, F. (2021). Examination of Community Sentiment Dynamics due to COVID-19 Pandemic: A Case Study from a State in Australia. *SN Computer Science*, 2(3). doi:https://doi.org/10.1007/s42979-021-00596-7.