WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

# Final Report

Evaluation of Twitter Data to find a connection between number of and public sentiment about Covid-19 Vaccines in US, UK and worldwide.


## Teammembers

Böhm Katharina

Jüngling Eva

Tavaszi Peter Gergely

Ulukhanov Hashym

## Instructors

Dr. Sabrina Kirrane

Astrid Krickl, MSc


Data Processing 2: Scalable Data Processing, Legal and Ethical Foundations of Data Science

# Table of Contents

# Project Overview

Our project report is structured in three sub-sections. We commence with the project overview where the project idea and its motivation are presented and continue with the project's objectives and how we intend to reach those. Furthermore, we discuss why it qualifies as a big data project and afterwards, discuss our data sources and our means to collect the data. Moreover, we present our proposed solution and the architecture, libraries and algorithms used.

The next section discusses legal and ethical issues. We describe the guidelines we followed, the legal and ethical challenges that came our way and how we overcame them.

At last, we review the challenges we encountered while working on the project and how we dealt with them. Furthermore, we outline the experience gained by each team member and conclude the report with recommendations for future work related to our project.

# Project Description and our Process

## The Project

Covid-19 has been a part of our lives since 2019/2020. Nation-wide lockdowns, social distancing and other restrictions have had a major impact on our society. For many people all over the world, Covid-19 vaccines are the end solution to this pandemic, however, not everyone is convinced by the vaccines. There are many different opinions and sentiments about this topic, all of which are expressed on social media. People who are yet undecided whether to get the vaccination shot or not might be influenced by this discourse of opinions and may reach their decision based on the information they receive, and the sentiments that are being conveyed on social media. Furthermore, there is a considerable amount of fake news and false information, like conspiration theories circulating on social media platforms, which might shape people's thinking and influence individuals' everyday decision-making.

To tackle the question whether a connection between the public reception and the number of vaccinations can be made, we will collect and compare tweets about different vaccine brands. The brands we look at specifically are Pfizer/BioNTech, Moderna, Johnson & Johnson's Janssen and AstraZeneca. We will analyse the tweets from the UK, the US, and on a worldwide level. The period we focus on starts with the 28th of June 2021 and ends with the 2nd of July 2021.

## Data Sources and methods used

### Twitter data: Sentiment Analysis

To gain an understanding of the public reception of the countries listed above, we collected tweets via Kafka-live streaming from the Twitter API data. We filtered for tweets in English language only, for the distinct locations (UK, US, worldwide) and for the different vaccine brands (Pfizer/BioNTech, Moderna, Johnson & Johnson's Janssen and AstraZeneca). Moreover, we conducted a sentiment analysis which is a natural language processing technique to determine the sentiment of the data collected. With sentiment analysis we detected the polarity and the subjectivity of the streamed tweets, and classified them as either negative, neutral, or positive.

We streamed the data daily at random times from the 28th of June 2021 until the 2nd of July 2021 until we had 'enough' tweets for each filter. Our filters were

- AstraZeneca in UK

- Pfizer/BioNTech in UK
- AstraZeneca in US
- Pfizer/BioNTech in US
- Pfizer Worldwide
- AstraZeneca Worldwide
- Johnson & Johnson's Janssen Worldwide
- Moderna Worldwide

As errors occurred in the streaming process and sentiment analysis sometimes failed to detect the sentiments of the tweets, we filtered for complete observations only.

Furthermore, to show what other topics people were concerned with during our selected time periods, we created word clouds displaying the most frequently used words related to each vaccine brand in the entire world.

### Vaccination Data

We retrieved our country-by-country data on Covid-19 vaccinations from the *Humanitarian Data Exchange*, which is "an open data sharing platform managed by the United Nations Office for the Coordination of Humanitarian Affairs." (HDX, 2021). The dataset was created and is currently maintained by *Our World in Data* and is being updated daily. It not only includes country-by-country data, but also data for subnational locations (e.g., England and Northern Ireland) and international aggregates (e.g., EU and Continents). The figures used in this dataset are retrieved from public official sources. The data we used from this dataset includes the location, the date, and the number of daily vaccinations. The latter records new doses administered in one day which however are seven-day-smoothed. As some countries do not publish data daily, it is assumed that the number of doses administered changes equally per day over any period in which no data was disclosed/not recorded. This in turn creates a series of daily numbers for which the average is taken over a rolling seven-day window. We extracted data only for the selected period.

### Machine Learning

We used various machine learning classifiers (Logistic Regression, Cross Validation, Naïve Bayes, Random Forest, Logistic Regression with TF-IDF Features) to predict to which of the three sentiment classes (positive, negative, neutral) the tweets belong. We used accuracy to compare these models. We also implemented linear regression models to see if average tweet sentiment has a significant effect on daily vaccination numbers across the observed regions and countries. For assessing the performance of the linear regression models, we used the Coefficient of Determination and Root Mean Squared Error. The exact implementation and the performance of these models will be discussed in the following sections of the report.

### Working with Big Data

Our project qualifies as Big Data project as it fulfils the five Vs - Volume, velocity, variety, veracity, and value – which is a concept often used to classify Big Data.

1. Volume: Twitter has 353 million monthly active users and 192 million daily active users (Dean, 2021). Thus, a vast number of tweets is posted daily, and big volumes of data are being created. Our code needed to sort through this data, detect the desired tweets and classify

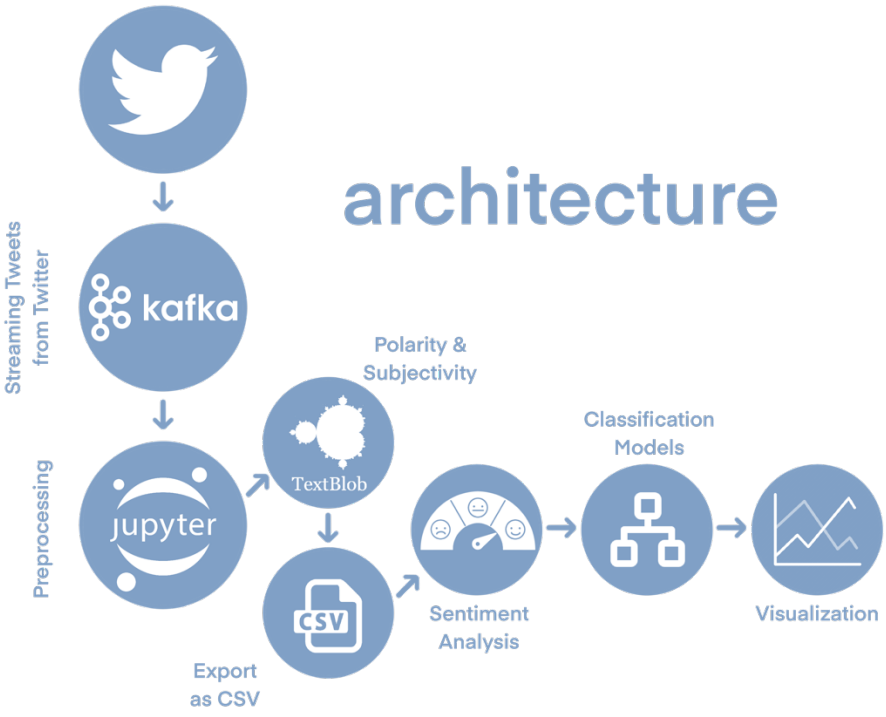
them for our analysis. Furthermore, the number of people being vaccinated increases worldwide daily, thus, we deal with a large volume of vaccination data.

2. Velocity: New tweets are being posted every second and call for high-speed processing. We streamed this data close to real-time through Apache Kafka.
3. Variety: We qualify tweets to be semi-structured as the tweet itself is structured but is made of various components such as the text of the tweet on which our sentiment analysis is conducted, it is unstructured. Furthermore, we combined the Twitter data with vaccination data.
4. Veracity: The veracity of Twitter is hard to assess as fake news can be easily created and distributed which impedes a true analysis of the public reception. Our vaccination data is based on officially published sources. Thus, it is available, and we trust the data to be authentic.
5. Value: Our project provides value as it can help to understand whether a connection exists between the way the population evaluates the vaccination and the number of people getting vaccinated. Since Covid-19 is a polarising topic, more insights into these aspects might aid in assessing the overall situation

## Description of the project solution:

### Pipeline Architecture:

We created a data architecture diagram to broadly illustrate the steps we took while solving the task at hand.




As the very first step, we used our developer accounts with the corresponding tokens and credentials to connect to the Twitter API and streamed tweets about the different COVID-19 vaccinations into Kafka-Python. In the second step, we downloaded the tweets as CSV files into our Jupyter environment. We repeated this procedure daily in the period ranging from the 28th of June 2021 until the 2nd of July 2021. One part of these CSV files was loaded into

and aggregated in the "Sentiment_Pipeline" notebook where we trained and fitted several supervised learning models on them to predict which sentiment classes they belonged to. The "worldwide_comparison" notebook was created to compare worldwide tweets about four different vaccination brands: Pfizer/BioNTech, AstraZeneca, Moderna, and Johnson & Johnson's Janssen. The "UK_US_comparison" notebook was created to compare UK tweets about AstraZeneca and Pfizer with US tweets about the same two vaccines. In these two notebooks we plotted the evolution of average sentiment polarity and trained linear regression models to predict the effect of tweet sentiments on vaccination rates in the distinct locations.

To analyse the evolution of vaccination rates in the U.K., the U.S., and on a worldwide level, we downloaded a country-by-country COVID-19 vaccination dataset from humdata.org (https://data.humdata.org/dataset/covid-19-vaccinations). The analysis, visualizations and insights gained from the collected social media data and the daily vaccination rates are presented on the following pages of this report.

### Streaming tweets from Twitter:

The producer and consumer notebooks that we used to stream the tweets from Twitter and load them into CSV-files were based on the course material we received in one of the lab sessions. We modified those notebooks to suit our specific needs.

In the first part of this project, we needed to generate our own data from Twitter so that we could conduct real-time public sentiment analysis on COVID-19 vaccines. We decided to set up several notebooks for the streaming process. Initially, we set up 24 different notebooks that we distributed equally between the four members of our group. The first six notebooks served the purpose of streaming tweets about the AstraZeneca vaccine. The second set of notebooks was created for streaming tweets about the Johnson & Johnson's Janssen vaccine, the third about Moderna, and the fourth about the Pfizer/BioNTech vaccine. We created one producer and one consumer notebook each for tweets coming in from the U.K., from the U.S., and from the entire world (no country filtering was adopted here) for each COVID vaccine brand.

We chose these four vaccination brands because they were the most popular and successful types of vaccines against the COVID-19 virus sold and used in the U.K. and the U.S. at the time when we were working on this project (June, July 2021). We could exploit the available storage space of a total of 2GB per person the most efficient way with each of us streaming one of the four types of vaccines. Each member of the group streamed around 50 to 60 different tweets daily for each geographical region and corresponding vaccine brand. We wanted to limit ourselves to a small but efficient number of daily tweets so that we would not run out of storage capacity.

We used Apache Spark Structured Streaming to work with streaming data. The streaming process can be described shortly the following way: first we imported the necessary libraries and packages (*tweepy*: for accessing the Twitter API and building the pipeline, *Kafka-python*: for loading the tweets into a Kafka topic, *JSON module*: for being able to work with the raw tweets that we received in the form of JSON objects). After that, we authenticated ourselves

with the credentials that we generated through our Twitter developer accounts. With the help of these tokens, we connected directly to the Twitter API.

In the next step, the StreamListener class of the producer continuously sent the tweets into a Kafka topic. The tweets were stored in raw JSON format in the Kafka topic. We set up the tracker to follow tweets that contain the relevant keywords about the corresponding vaccinations, and we only streamed tweets in English.

### Processing of Twitter Data:

Once the streaming process was set up and running, we moved on to our consumer notebooks, where we first needed to create the connection between Kafka Python and Apache Spark.

Each consumer notebook received the tweets via subscription to the Kafka topic that was created in the producer notebook. We imported the *PySpark* library to preprocess the Twitter data, *TextBlob* package from the *textblob* library to extract sentiment polarity and subjectivity scores from the streamed tweets' main text with the help of user-defined functions. We built an empty SparkSession and connected it to the Kafka topic, which we defined in the producer notebook. Furthermore, we transformed the sentiment polarity scores (between -1 and 1) into three classes of strings: “negative” if lower than 0, “positive” if higher than 0, and “neutral” if exactly 0.

[21]:
```
# text classification

# Define methods from TextBlob
def polarity_detection(text):
    return TextBlob(text).sentiment.polarity

def subjectivity_detection(text):
    return TextBlob(text).sentiment.subjectivity

def sentiment_detection(value):
    if value < 0:
        return 'Negative'
    elif value > 0:
        return 'Positive'
    else:
        return 'Neutral'

# polarity detection
# Define as user defined fuction to embed method in the spark environment
polarity_detection_udf = udf(polarity_detection, StringType())

# subjectivity detection
# Define as user defined fuction to embed method in the spark environment
subjectivity_detection_udf = udf(subjectivity_detection, StringType())

# sentiment detection
# Define as user defined fuction to embed method in the spark environment
sentiment_detection_udf = udf(sentiment_detection, StringType())
```

We performed some essential data processing steps in the consumer notebook so that the tweets were loaded into the CSV-files in perfect shape and ready for further analysis. We extracted the most essential information from each tweet that we received: The text of the tweet, the date and timestamp when it was created, and the location where it was created. We also appended the corresponding polarity and subjectivity scores as well as the textual sentiment class to the dataframe.

Since we needed the text of the tweets in the cleanest and best shape possible, we implemented some more sophisticated pre-processing steps on the text column of the table. With the help of regular expressions, we removed the links, hashtags, mentioned usernames, the “RT” (which indicates retweets), and the “:” character.

```
[ ]: try:
    # Cast the data into a json
    tweet_df_string = tweet_df.selectExpr("CAST(value AS STRING) as json_data")

    # extract the tweet and user info
    text_user = tweet_df_string.select(json_tuple('json_data', 'created_at','text', 'user').alias('created_at', 'text',
                                                                                                   'json_user'))

    # extract screen_name and location from user info
    text_user_info = text_user.select('text', 'created_at', json_tuple('json_user', 'location').alias('location'))

    # preprocessing
    text_user_info = text_user_info.na.replace('', 'None')
    text_user_info = text_user_info.na.drop()

    text_user_info = text_user_info.withColumn('text', F.regexp_replace('text', r'http\S+', ''))
    text_user_info = text_user_info.withColumn('text', F.regexp_replace('text', '@\w+', ''))
    text_user_info = text_user_info.withColumn('text', F.regexp_replace('text', '#', ''))
    text_user_info = text_user_info.withColumn('text', F.regexp_replace('text', 'RT', ''))
    text_user_info = text_user_info.withColumn('text', F.regexp_replace('text', ':', ''))
```

Finally, the tweets were compiled into dataframes and written into CSV-files every 60 seconds. The individual CSV files were first stored in the Jupyter environment and later we combined them in a separate notebook (“file_combination”).

## Sentiment Analysis Pipeline:

At this stage, we have collected a large amount of data from Twitter with the help of the previous notebooks and we could go on to conduct an in-depth sentiment analysis of the tweets. We started by loading the combined worldwide Twitter CSV-file into the “Sentiment_Pipeline” notebook.

We started by importing the *findspark* and *nltk* modules, building a SparkSession and loading the merged CSV-file of all worldwide Twitter files about all vaccination types into the notebook as a Pyspark dataframe. First, we inspected the basic characteristics and parameters of the table e.g., size, and potential missing values. We removed all missing values, converted the dates column to datatype date and ordered the tweets by date.

```
[7]: try:
    # Get size and shape of DF
    print((worldwide_combined_df.count(), len(worldwide_combined_df.columns)))
    # Display head of DF (we see some missing values)
    worldwide_combined_df.show(10)
except:
    print("Unexpected error:", sys.exc_info()[0])
```

```
(3606, 6)
+--------------------+--------------------+--------------------+--------+------------+---------+
|                text|                date|            location|polarity|subjectivity|sentiment|
+--------------------+--------------------+--------------------+--------+------------+---------+
|The Pfizer-BioNTe...|Mon Jun 28 14:43:...|Johannesburg, Sou...|     0.5|         0.5| Positive|
|    this is true.|                null|                null|    null|        null|     null|
|Pfizer and Modern...|Mon Jun 28 14:43:...|       United States|    0.35|        0.65| Positive|
|The Pfizer-BioNTe...|Mon Jun 28 14:43:...|              Mexico|     0.5|         0.5| Positive|
|Great news – Pfiz...|Mon Jun 28 14:43:...|               World|     0.8|        0.75| Positive|
|The Pfizer-BioNTe...|Mon Jun 28 14:43:...|Südlich von Münch...|     0.5|         0.5| Positive|
|I should be somew...|Mon Jun 28 14:43:...|Region Of Waterlo...|    -0.4|         0.8| Negative|
|You truly love to...|Mon Jun 28 14:44:...|             Toronto|     0.5|         0.6| Positive|
|You truly love to...|Mon Jun 28 14:44:...|        Brooklyn, NY|     0.5|         0.6| Positive|
|The Pfizer-BioNTe...|Mon Jun 28 14:44:...|    Philadelphia, PA|     0.5|         0.5| Positive|
+--------------------+--------------------+--------------------+--------+------------+---------+
only showing top 10 rows
```

[8]:
```
try:
    # drop all tweets with missing values
    worldwide_combined_df = worldwide_combined_df.na.drop("any")
    print((worldwide_combined_df.count(), len(worldwide_combined_df.columns)))
    worldwide_combined_df.show(10)
except:
    print("Unexpected error:", sys.exc_info()[0])
```

```
(2395, 6)
+--------------------+--------------------+--------------------+--------+------------+---------+
|                text|                date|            location|polarity|subjectivity|sentiment|
+--------------------+--------------------+--------------------+--------+------------+---------+
|The Pfizer-BioNTe...|Mon Jun 28 14:43:...|Johannesburg, Sou...|     0.5|         0.5| Positive|
|Pfizer and Modern...|Mon Jun 28 14:43:...|       United States|    0.35|        0.65| Positive|
|The Pfizer-BioNTe...|Mon Jun 28 14:43:...|              Mexico|     0.5|         0.5| Positive|
|Great news – Pfiz...|Mon Jun 28 14:43:...|               World|     0.8|        0.75| Positive|
|The Pfizer-BioNTe...|Mon Jun 28 14:43:...|Südlich von Münch...|     0.5|         0.5| Positive|
|I should be somew...|Mon Jun 28 14:43:...|Region Of Waterlo...|    -0.4|         0.8| Negative|
|You truly love to...|Mon Jun 28 14:44:...|             Toronto|     0.5|         0.6| Positive|
|You truly love to...|Mon Jun 28 14:44:...|        Brooklyn, NY|     0.5|         0.6| Positive|
|The Pfizer-BioNTe...|Mon Jun 28 14:44:...|    Philadelphia, PA|     0.5|         0.5| Positive|
|The Pfizer-BioNTe...|Mon Jun 28 14:44:...|          Lima, Perú|     0.5|         0.5| Positive|
+--------------------+--------------------+--------------------+--------+------------+---------+
only showing top 10 rows
```

[9]:
```
'''
https://stackoverflow.com/questions/68239001/pyspark-non-matching-values-in-date-time-column
'''
try:
    import pyspark.sql.functions as F
    # conveting to dateformat and stripping away timestamp according to stackoverflow response
    worldwide_combined_df = worldwide_combined_df.withColumn("date", F.to_date(F.substring("date",5,100),"MMM dd HH:mm:ss xx
    worldwide_combined_df.show(10)
except:
    print("Unexpected error:", sys.exc_info()[0])
```

```
+--------------------+----------+--------------------+--------+------------+---------+
|                text|      date|            location|polarity|subjectivity|sentiment|
+--------------------+----------+--------------------+--------+------------+---------+
|The Pfizer-BioNTe...|2021-06-28|Johannesburg, Sou...|     0.5|         0.5| Positive|
|Pfizer and Modern...|2021-06-28|       United States|    0.35|        0.65| Positive|
|The Pfizer-BioNTe...|2021-06-28|              Mexico|     0.5|         0.5| Positive|
|Great news – Pfiz...|2021-06-28|               World|     0.8|        0.75| Positive|
|The Pfizer-BioNTe...|2021-06-28|Südlich von Münch...|     0.5|         0.5| Positive|
|I should be somew...|2021-06-28|Region Of Waterlo...|    -0.4|         0.8| Negative|
|You truly love to...|2021-06-28|             Toronto|     0.5|         0.6| Positive|
|You truly love to...|2021-06-28|        Brooklyn, NY|     0.5|         0.6| Positive|
|The Pfizer-BioNTe...|2021-06-28|    Philadelphia, PA|     0.5|         0.5| Positive|
|The Pfizer-BioNTe...|2021-06-28|          Lima, Perú|     0.5|         0.5| Positive|
+--------------------+----------+--------------------+--------+------------+---------+
only showing top 10 rows
```

[10]:
```
try:
    # drop any remaining tweets with missing values and order the datframe by date
    worldwide_combined_df = worldwide_combined_df.na.drop("any")
    worldwide_combined_df = worldwide_combined_df.orderBy("date")
    print((worldwide_combined_df.count(), len(worldwide_combined_df.columns)))
    worldwide_combined_df.show(10)
except:
    print("Unexpected error:", sys.exc_info()[0])
```

```
(2329, 6)
+--------------------+----------+--------------------+--------+------------+---------+
|                text|      date|            location|polarity|subjectivity|sentiment|
+--------------------+----------+--------------------+--------+------------+---------+
|The Pfizer-BioNTe...|2021-06-28|Johannesburg, Sou...|     0.5|         0.5| Positive|
|Pfizer and Modern...|2021-06-28|       United States|    0.35|        0.65| Positive|
|The Pfizer-BioNTe...|2021-06-28|              Mexico|     0.5|         0.5| Positive|
|Great news – Pfiz...|2021-06-28|               World|     0.8|        0.75| Positive|
|The Pfizer-BioNTe...|2021-06-28|Südlich von Münch...|     0.5|         0.5| Positive|
|I should be somew...|2021-06-28|Region Of Waterlo...|    -0.4|         0.8| Negative|
|You truly love to...|2021-06-28|             Toronto|     0.5|         0.6| Positive|
|You truly love to...|2021-06-28|        Brooklyn, NY|     0.5|         0.6| Positive|
|The Pfizer-BioNTe...|2021-06-28|    Philadelphia, PA|     0.5|         0.5| Positive|
|The Pfizer-BioNTe...|2021-06-28|          Lima, Perú|     0.5|         0.5| Positive|
+--------------------+----------+--------------------+--------+------------+---------+
only showing top 10 rows
```

The original dataframes that we streamed from Twitter contained the columns 'text', 'created_at', 'location', 'polarity', subjectivity' and 'sentiment'. We created a new column 'tweet_length' by measuring the length of the 'text' column for each individual tweet. We could gain further insights about the data by printing out the length of the longest and the shortest tweet, respectively.

The next significant step in the sentiment analysis process was building the sentiment analysis pipeline. When dealing with text processing with to goal of obtaining deeper insights via machine learning algorithms, it is recommended to use the MlLib Pipeline function. The text processing workflow usually consists of several PipelineStages that must be run in a specific order on the original Twitter dataframe. The PipelineStages consist of a series of Transformers and Estimators, which reshape the entire dataframe systematically. Each stage transforms the original dataframe in the specified manner and passes the updated dataframe to the next stage. Our issue was that Spark's built-in Pipeline function only accepted built-in transformers and estimators and therefore it was only suited for simpler data workflows. However, we intended to include some more sophisticated steps in the analysis by the means of user-defined functions, since we have not found any appropriate built-in versions of these functions in Spark's SQL function set. Although we did not use it directly, the built-in MlLib Pipeline was our methodological guideline for the implementation of our homemade sentiment pipeline.

Our self-made machine-learning pipeline consisted of a built-in RegexTokenizer that took the 'text' column as input and split it into an array of words. Further, it contained a StopWordsRemover that took in the list of words created by the tokenizer, removed redundant words that did not convey additional information for our sentiment analysis, like "does", "a", "an", "the", "at", "by", "for", etc. and put out the column "filtered_words".

[14]:
```
'''
https://spark.apache.org/docs/latest/ml-features.html
'''
try:
    # stop words remover: removes words that do not convey additional info
    add_stopwords = ["i", "me", "my", "myself", "we", "our", "ours", "ourselves",
                     "you", "your", "yours", "yourself", "yourselves",
                     "he", "him", "his", "himself", "she", "her", "hers", "herself",
                     "it", "its", "itself", "they", "them", "their", "theirs", "themselves",
                     "what", "which", "who", "whom", "this", "that", "these", "those",
                     "am", "is", "are", "was", "were", "be", "been", "being",
                     "have", "has", "had", "having", "do", "does", "did", "doing",
                     "a", "an", "the", "and", "but", "if", "or", "because", "as",
                     "until", "while", "of", "at", "by", "for", "with", "about", "against", "between",
                     "into", "through", "during", "before", "after", "above", "below",
                     "to", "from", "up", "down", "in", "out", "on", "off", "over", "under",
                     "again", "further", "then", "once", "here", "there", "when", "where", "why", "how",
                     "all", "any", "both", "each", "few", "more", "most", "other", "some",
                     "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very",
                     "s", "t", "can", "will", "just", "don", "should", "now"]
    remover = StopWordsRemover(inputCol="words", outputCol="filtered_words").setStopWords(add_stopwords)
    filteredData = remover.transform(wordsData)
    filteredData.select("text", "words", "filtered_words").show(truncate=True)
except:
    print("Unexpected error:", sys.exc_info()[0])
```

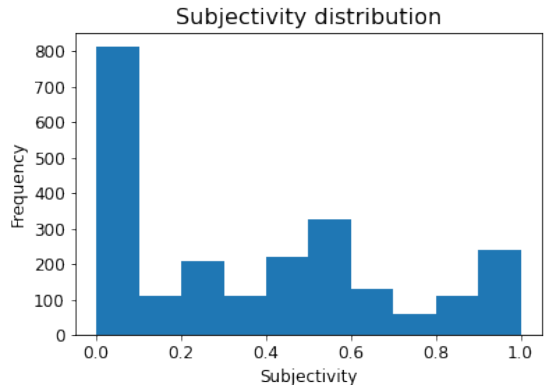
We also created a series of user-defined functions to analyse the dataframe further. At the next stage, we created the removePunctuation udf, which took the "filtered_words" column as input, removed all punctuation and other irrelevant grammatical attributes from it, and

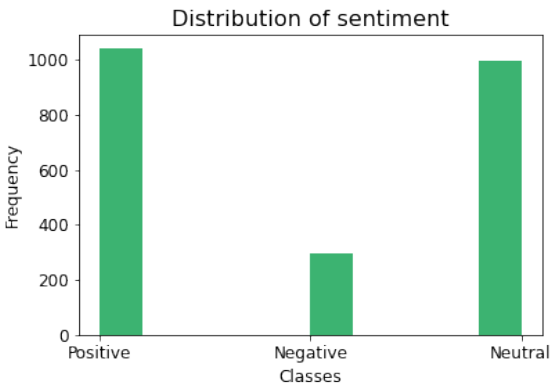
put out the “punc_cleaned” column. After having dealt with punctuation, the next udf, a lemmatizer function took the “filtered_words” column of the pipeline dataframe as input and transformed the words into their basic grammatical form. This process works by determining the correct ‘part-of-speech’ tag and mapping it to the corresponding word. Examples for the transformations performed by the lemmatizer functions would be: are -> *be,* removed -> *remove,* hanging -> *hang,* vectors -> *vector*

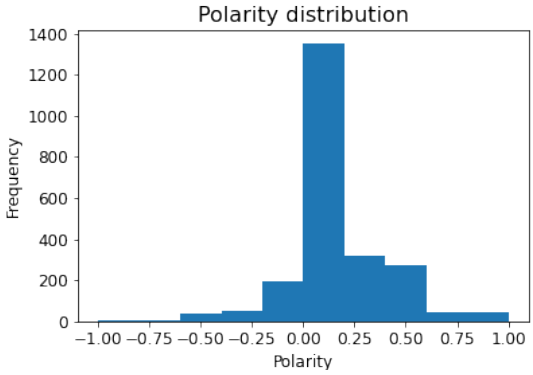
The last udf was based on a much simpler methodology. It counted the tokens in the “lemmatized_words” column and returned length of the lemmatized version of each tweet. We additionally created a separate column called “ngrams”, which contained bigrams that were created based on the “lemmatized_words” column with the help of the built-in NGram MlLib function.

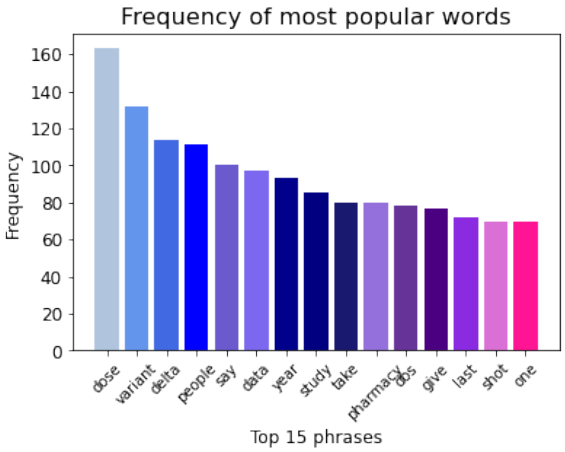
[15]:
```
'''
https://stackoverflow.com/questions/58038919/removing-punctuation-in-spark-dataframe
'''
try:
    # punctuation remover: removes commas, dots and grammatical signs
    def removePunctuation(column):
        '''Removes punctuation, changes to lower case, strips leading and trailing spaces and splits on separator string.
            Note:
            Only spaces, letters, and numbers should be retained.
            Args:
            column (Column): filtered_words column from previous step.
            Returns:
            Column: column named 'punc_cleaned' with clean-up operations applied.
        '''
        #return lower(trim(regexp_replace(column,'\\p{Punct}','')))
        return split(trim(lower(regexp_replace(concat_ws(" ", column),'\\p{Punct}', ''))), " ")

    #PunctRemover = udf(removePunctuation, ArrayType(StringType()))

    filteredData = filteredData.withColumn("punc_cleaned", removePunctuation(filteredData.filtered_words))
    filteredData.select("text", "words", "filtered_words", "punc_cleaned").show(5, vertical=True, truncate=False)
except:
    print("Unexpected error:", sys.exc_info()[0])
```

Most importantly, the final stage of our pipeline featured a CountVectorizer, which transformed the array of words into a term frequency vector. Creating the term frequency vector is an essential part of the machine learning process because the computer algorithms can always work more efficiently with numerical data than with raw textual data. The last element of the pipeline was the StringIndexer, which encoded the textual sentiment labels into numeric indices. These indices were labelled by class frequencies: 0.0 for the most frequent class, 1.0 for the second most frequent class and 2.0 for the least frequent class. We also created some visualizations based on the information we extracted from the tweets with the help of the pipeline.




## Classification models for Twitter Data:

After having analysed our data and having performed the various feature extraction techniques outlined above, we were ready to implement some of the commonly used text classification models. The various models and their evaluations helped us gain a broad overview of the underlying multi-class text classification problem and predict the sentiment classes of tweets.

We started by randomly splitting the dataframe into a 70% training and a 30% test set with a seed set for reproducibility. The first supervised machine-learning algorithm we trained was the multinomial Logistic Regression model. We first fitted it to the training dataset and afterwards fitted it on the test dataset using the *.tranform()* method to see how it performs on unseen data. For the model evaluation, we used Spark's MulticlassClassificationEvaluator, which assessed the accuracy of the predictions. Using the training summary of the Logistic Regression function, we were able to obtain further evaluation metrics like the objective per iteration. Furthermore, for multiclass Logistic Regression, we were also able to inspect metrics like true positive rate, false positive rate, recall, precision and F-measure by label. The performance of the logistic regression function was very poor with approximately 27% accuracy. We suspect that it predicted the labels so poorly because it is generally designed for binary classification problems and thus has significant troubles when working on a multiclass classification problem.

We repeated the same steps that we outlined above to fit a Logistic Regression model with 5-fold cross validation, a Naïve Bayes classifier, as well as a Random Forest classification model to see, which model performs best with our Twitter data. Logistic Regression with 5-

fold cross validation was our best classification model and had the highest accuracy with approximately 81%. Naïve Bayes was the second-best classification model with 74% accuracy, and it was followed by Random Forest with 51% accuracy.

Next, we built a new pipeline for Logistic Regression with Term Frequency – Inverse Document Frequency (TF-IDF). TF-IDF is a statistic that measures how important a word is to a document in a corpus. It is composed of two metrics, the term frequency i.e., how many times a word appears in a document and the inverse document frequency i.e., how frequently a word appears in the entire corpus of documents. Inverse document frequency can be calculated by dividing the total number of documents by the number of documents that contain a word and taking the logarithm of the whole term. Finally, we obtain the TF-IDF by multiplying these two metrics with each other. The higher the TF-IDF score of a word, the more significant that word is in a specific document. TF-IDF is a computer algorithm that helps us transform a text into a numerical vector, so that it is easier to handle for the machine learning algorithms, like classifiers. Unfortunately, the Logistic Regression model with TF-IDF Features performed as poorly as the standard Logistic Regression model in predicting sentiment classes with 27% accuracy.

## Visualizations:

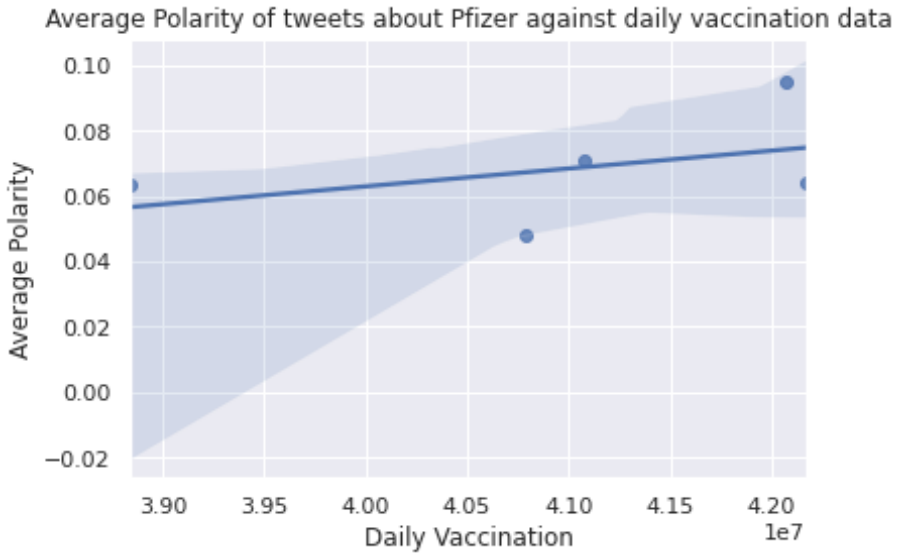
Lastly, we moved on to the two notebooks in which we presented our visualizations of the analysis of the Twitter data. In the first notebook called "worldwide_comparison", we started the visualization process by loading the aggregated the worldwide CSV-files separated by vaccination types. As a result, we got four tables each containing tweets about a specific vaccination brand. In the second notebook ("UK_US_comparison"), we aggregated all UK tweets about the Pfizer vaccine, all UK tweets about the AstraZeneca vaccine, all US tweets about the Pfizer vaccine, and all US tweets about the AstraZeneca vaccine into four different tables. The following steps were the same in both notebooks.

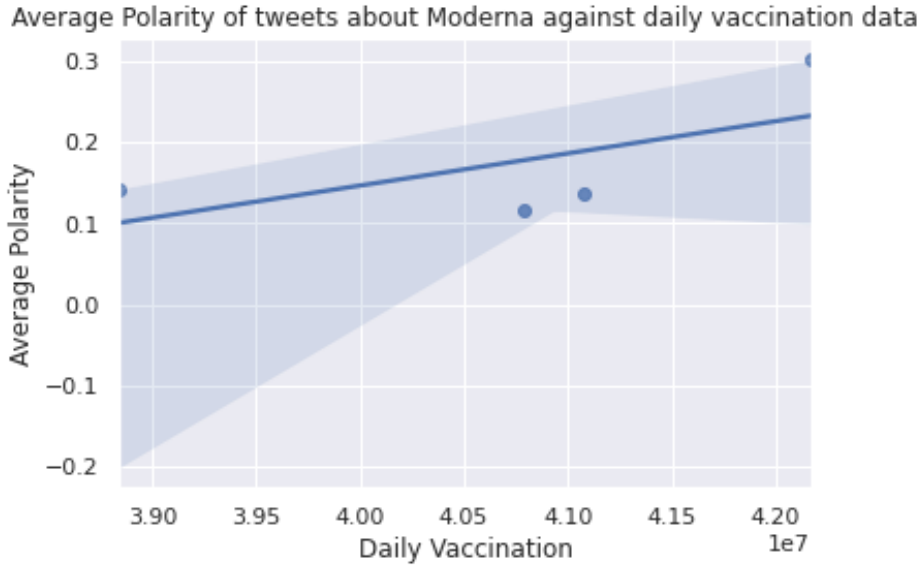
As a first step, we removed all missing values from the tables and converted the dates to datatype date and the polarity and subjectivity scores to floats. After that, we ordered the dataframes by date, grouped the tweets by days and took the average of the sentiment polarity and subjectivity scores. Hence, we received another four tables with daily average polarities and subjectivities.

In the next step, we loaded the vaccination data into the Jupyter notebook. We selected the "location", "date", and "daily_vaccinations" columns and filtered the table on the date column for the observation period between the 28$^{th}$ of June 2021 and the 2$^{nd}$ of July 2021. Finally, we joined the vaccination table with the table containing the daily average polarities and subjectivities.

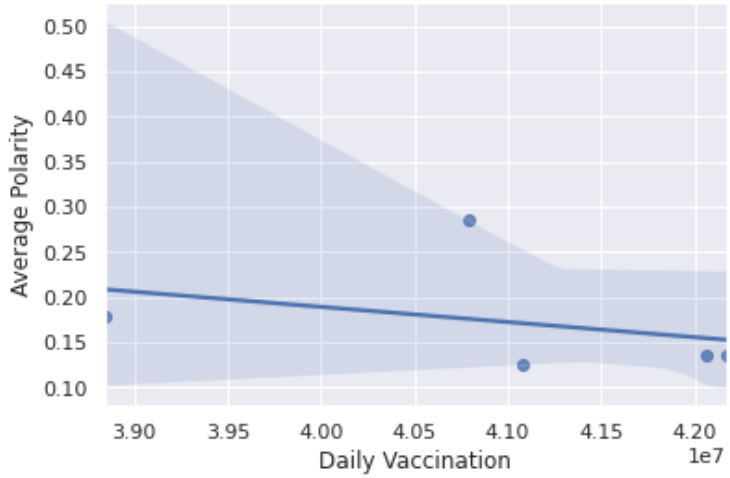
For the first visualization, we plotted the average polarity against the vaccination rates in a scatterplot. We repeated this graph for each different vaccination type we streamed tweets about in the first notebook. In the second notebook, we created four plots based on vaccination brand and geographical location (UK and US). It is difficult to draw an overarching conclusion from looking at these plots because we have only observed Twitter users' activity for five days and there is some variation in the resulting plots. These plots are meant to

visualize the correlation between the two variables. We interestingly observed a moderately large negative correlation coefficient for the Johnson & Johnson vaccine worldwide (-0.68) and a low negative coefficient for UK tweets about AstraZeneca (-0.02) and Pfizer (-0.12) between average sentiment polarity and daily vaccination numbers. In other words, the lower the Twitter sentiment polarity score the higher the number of daily vaccinated people. For the other three vaccination types on a worldwide scale, we observed positive correlation coefficients that can also be seen in these graphs. The US tweets about AstraZeneca and Pfizer follow this path and also result in positive coefficients.

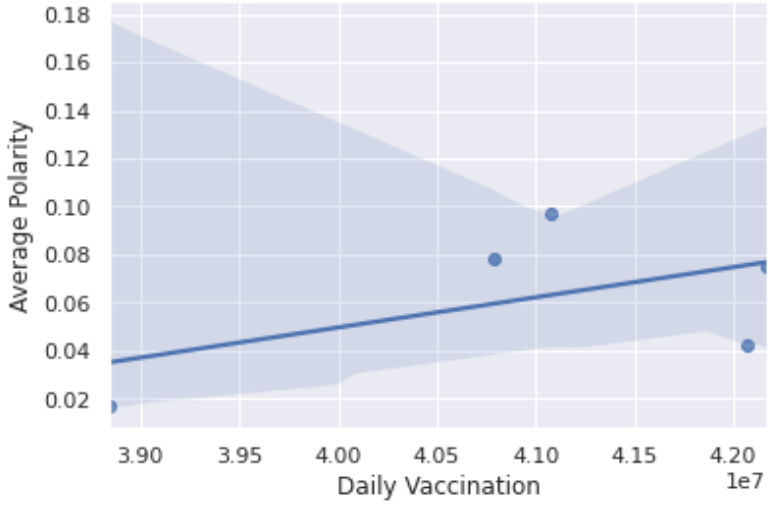


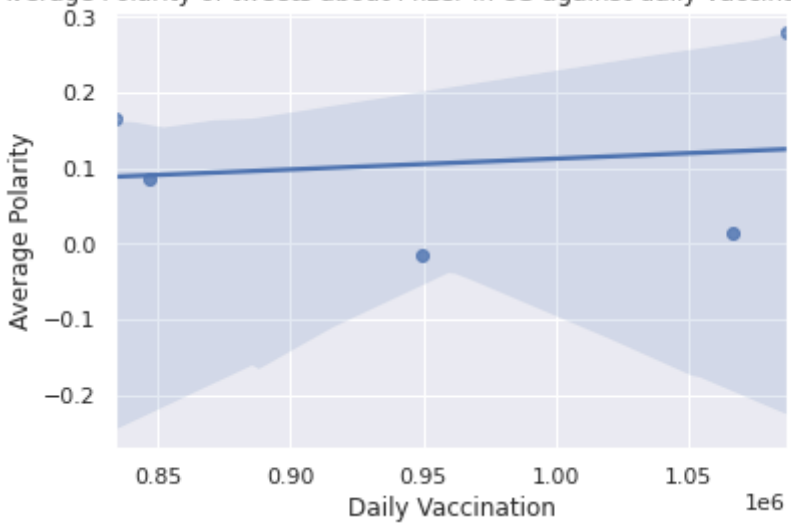
Average Polarity of tweets about Johnson & Johnson against daily vaccination data



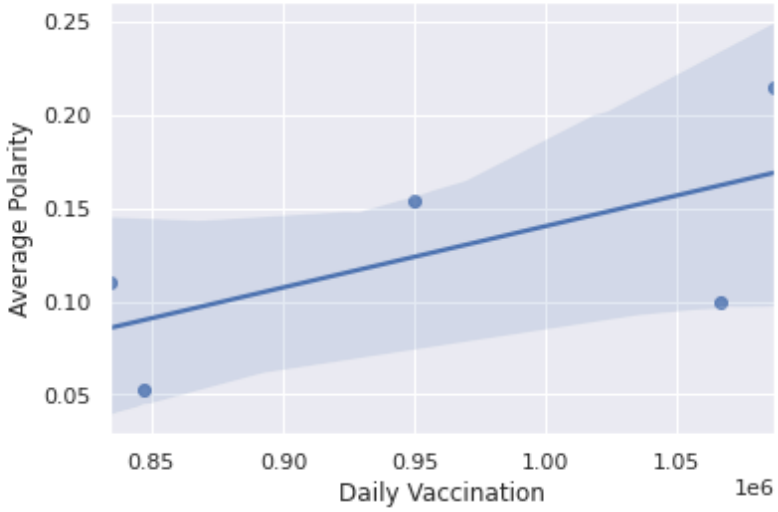

Average Polarity of tweets about AstraZeneca against daily vaccination data

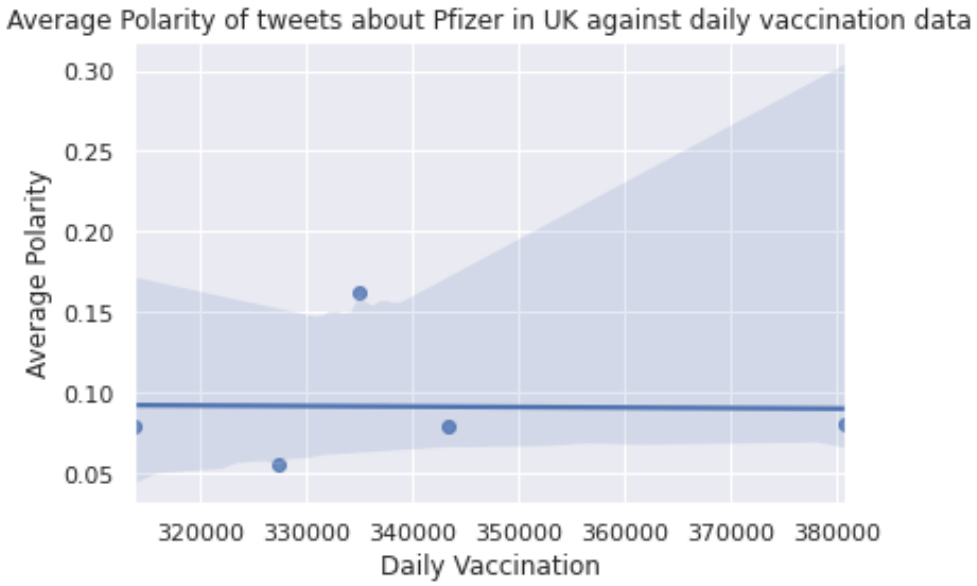


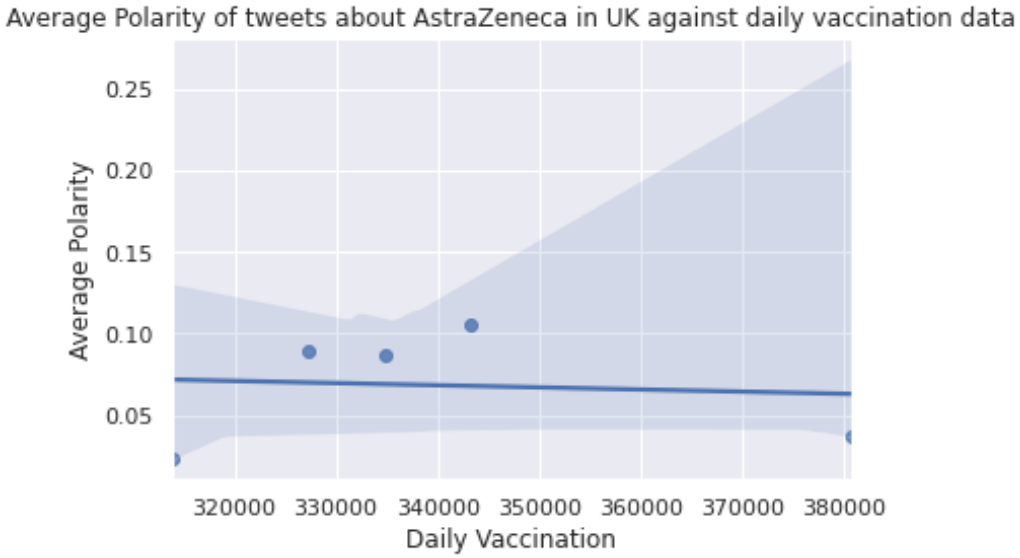
Average Polarity of tweets about Pfizer in US against daily vaccination data




Average Polarity of tweets about AstraZeneca in US against daily vaccination data

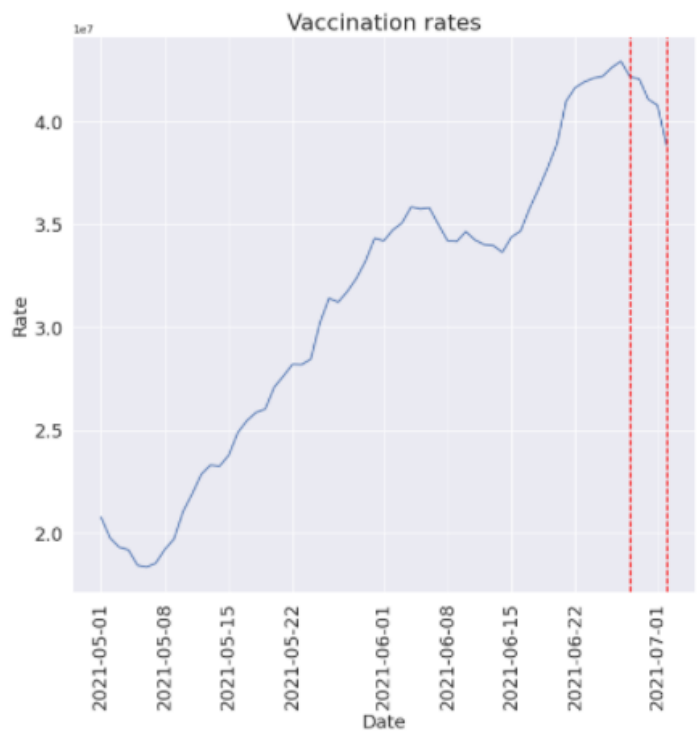


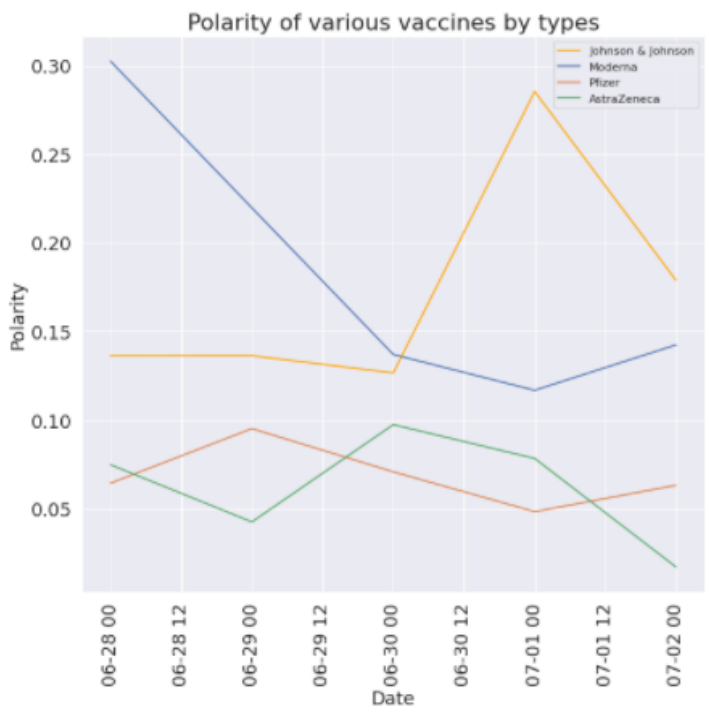


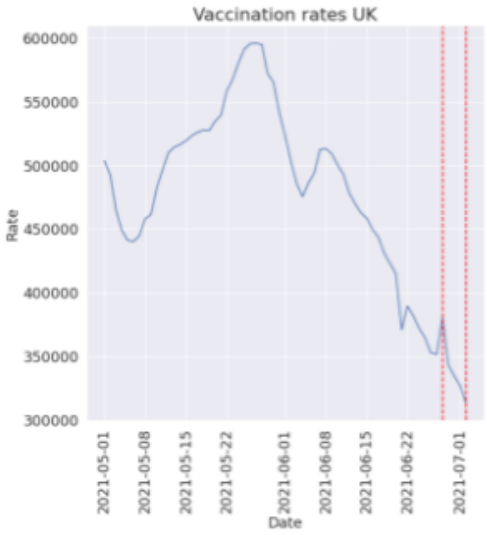




For the second visualization in the “worldwide_comparison” notebook, we joined the daily average polarity scores for each vaccination brand in one table and plotted them as a time series with dates on the x-axis. Next to this graph, we plotted the evolution of daily vaccinations worldwide and we could clearly observe a trend of decreasing daily vaccination numbers while the average sentiment polarities of tweets tended to slightly decrease as well. In the UK_US_comparison notebook we could discover a similar decreasing trend of daily vaccinations in the observation period for the UK whilst the average sentiment polarity scores tended to decline somewhat. We can also see from the graph that daily vaccinations in the US started to recover from a long decline in the observation period and at the same time, we saw an increase in average polarity. We cannot make a clear conclusion whether a decline in daily vaccinations is attributable to a short-term slowdown in the manufacturing process of the vaccines, some supply chain issues or perhaps even a general negative shift in the attitude of people towards vaccinating themselves. However, we think that this trend is definitely worth further investigation and research. Finding the reason for this negative trend could

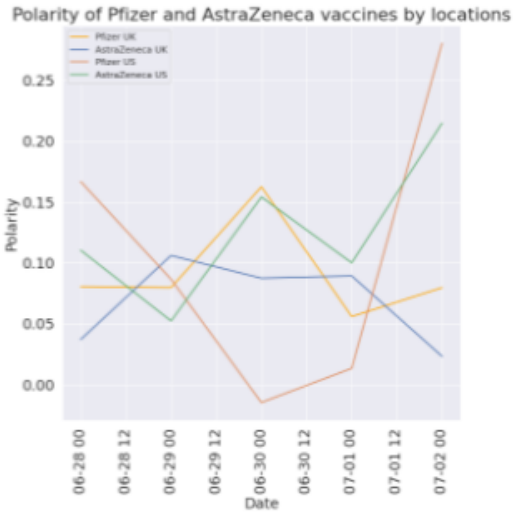
potentially support the operation of national health care systems all over the world and boost the sales numbers of the vaccination companies.








## Linear Regression:

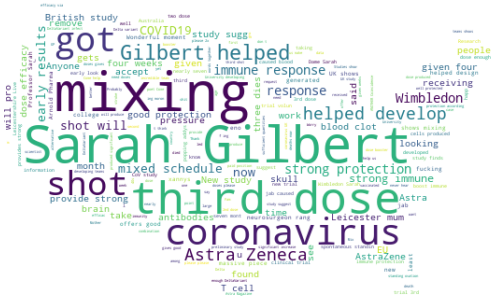
Lastly, we trained linear regression models on the dataframes containing daily vaccination rates and daily average polarity and subjectivity scores to measure the effect of sentiment polarity of tweets on daily vaccination numbers. First, we measured the correlation between average polarity and daily vaccination numbers as well as between average subjectivity and daily vaccination numbers. The correlation coefficients were mostly in the lower positive region, except for the Johnson & Johnson vaccine tweets, which have produced negative coefficients. In the next step, we used a VectorAssembler to transform the average polarity scores into features for the linear regression function. We first tried to split the dataframe

into a training and a test set to assess the model performance on unseen data, but that unfortunately left us with unacceptable results because the linear regression line was fitted to only three (training set) and two (test set) data points, respectively. Therefore, we finally trained the linear regression function on the table containing all five daily average polarities and subjectivities as well as the five daily vaccination numbers. We evaluated our linear regression models based on two commonly used metrics: the Coefficient of Determination (R2) and Root Mean Squared Error (RMSE). R-squared is the proportion of the variance for a dependent variable, which is explained by one or more independent variable in the model. R2 was 11.8% for worldwide Johnson & Johnson tweets, 40.9% for worldwide Moderna tweets, 18.5% for worldwide Pfizer tweets, and 27.8% for worldwide AstraZeneca tweets. In the "UK_US_comparison" notebook, we got an R2 of 0.04% for UK Pfizer tweets, 0.8% for UK AstraZeneca tweets, 2% for US Pfizer tweets, and 39.4% for US AstraZeneca tweets. RMSE was mainly in an acceptable range i.e., small enough compared to the variation of the actual dependent variable for all models except for US AstraZeneca and US Pfizer tweets. Even though we could not test the models on unseen data due to a lack of observations, our models performed fairly well on the small training data set that we were able to collect from the 28th of June until the 2nd of July 2021.

## Word Clouds:

As a last step we investigated the combined worldwide dataframes in the "worldwide_comparison" notebook and the combined UK and US dataframes in the "UK_US_comparison" notebook with the help of word clouds. We started by transforming the 'text' column of each table into a single combined string using the *.collect()* method. Once we had all the textual data in one string, we could use it to generate a word cloud from it. We imported the WordCloud package, added additional stop words to increase the cloud's educational power and put impressive masks on the clouds to make them look more appealing. The word cloud was an efficient tool to get an initial idea of which topics were discussed at that time in relation to COVID-19 vaccinations and which keywords appeared most frequently in conversations and debates. The larger the font size of a word in the cloud the more important it was in daily conversation.

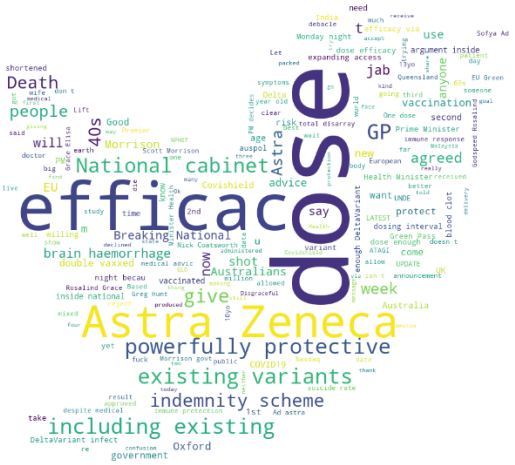

Words in US Pfizer tweets



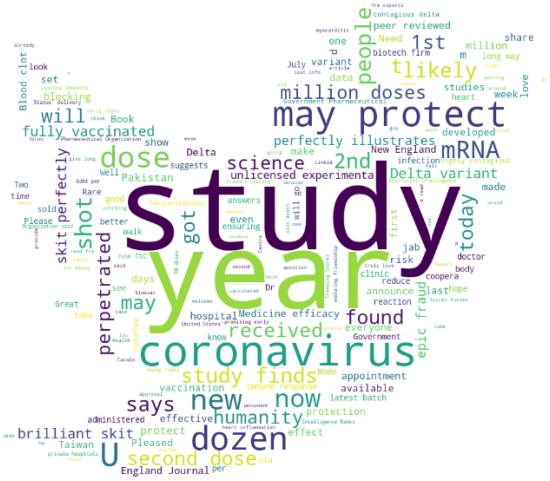

Words in US AstraZeneca tweets


Words in UK Pfizer tweets

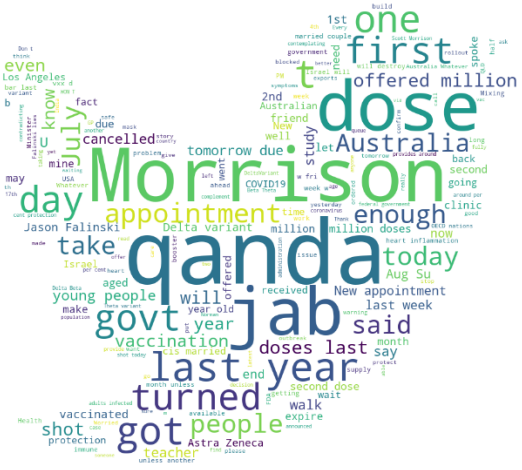

Words in UK AstraZeneca tweets



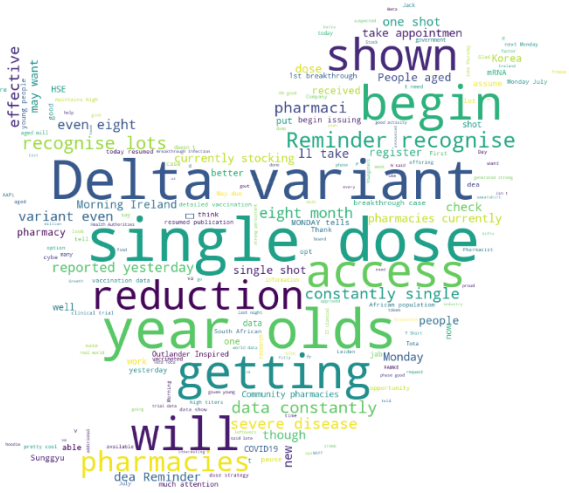

Words in worldwide AstraZeneca tweets




Words in worldwide Moderna tweets




Words in worldwide Pfizer tweets




Words in worldwide Johnson & Johnson tweets

## Interpretation and Conclusion of our solution

The way we approached our solution left us with many different outcomes to consider when drawing a conclusion – it actually leads to different conclusions across the observed vaccines and countries. Additionally, it is crucial to observe the data and the resulting analysis in connection with current events that influence the sentiment of Twitter users.

Especially the word clouds allow informative insights into what topics are especially connected with the respective vaccine or country. It is salient, that the words like “new”, “variant” or “delta” have been of great concern across all the vaccines. We conclude that the rise of the delta variant has caused this spike of interest, as it has (at least in the UK) become the dominant strain of coronavirus. The vaccines, as most important way to address the concern, are now investigated as to whether they will be able to provide the same amount of protection especially given the fact that the new variants are said to lead to more hospitalizations than previous ones (O’Dowd, 2021). The single dose of the Johnson & Johnson vaccine also is of interest for the users.

When taking a closer look at the AstraZeneca word clouds, the initial positive impression derived from the sentiment of the tweets is overshadowed by negative associations such as the connection to blood clots, brain haemorrhage or death which are the most prominent with this type of vaccine and hardly show up for any of the others. The third dose is also part of the discussion, but it is not intuitive to determine whether the connotation there is positive or negative, as is Sarah Gilbert as one of the people responsible for the development of this vaccine.

Now that we have identified certain concerns of Twitter users in connection with the vaccines, we can have a closer look at how they may have impacted the number of vaccines across countries and the entire world.

Our initial idea was to assess, whether a connection can be made between the number of vaccinations and the sentiment. Does a positive correlation mean that more people will get their vaccine if they read positive reviews on Twitter? Or does an increasing number of people receive their vaccine and share their opinion afterwards? This is difficult to determine but when looking at our results, we would assume the latter to be the case. When looking at the big picture that the worldwide analysis shows, we can overall see a positive connection between the numbers and the polarity – this would lead us to the assumption that people who received their vaccine are pleased with it. From personal

experience, the feeling of safety or relief could be a contributing factor here. It is however prominent, that the polarity of the Johnson & Johnson vaccine has declined compared to the others and does not seem to enjoy as good of a reputation. This could be traced back to the worries of users that the single dose of the vaccine might not protect as well against the new variant as the others do. These concerns have been addressed by various newspapers or magazines such as the New York Magazine which outline for example the possibility that a follow-up with an mRNA vaccine might be needed to obtain the optimal protection (Rosa-Aquino, 2021). Thus, we would explain the drop in polarity with the rise of the delta variant that causes users to worry about the effectiveness of this vaccine.

When looking at the US, it is surprising to see, that the polarity of AstraZeneca rose significantly over the 5 days of streaming while it almost stayed the same for Pfizer, which was still positive on average. The situation is entirely different for the UK where both vaccines are in decline in terms of polarity. Given the current numbers of the two countries, those developments make sense. While the US is currently experiencing stable development, the UK struggles with rising numbers and the possibility of upcoming lockdowns. The 7-day average over the past week (29$^{th}$ of June to 5$^{th}$ of July) has showed a 10 times higher percentage of the population being newly infected in the UK than in the US. Given the similar vaccination rates of 47.9% in the US and 50.6% in the UK, it is no surprise that the sentiment towards the vaccine differs in the two countries. When having similar rates of vaccination, the general expectation would be, that the numbers of new infections are similar too. Our interpretation is that the sentiment towards the vaccines is less positive in the UK because they were supposed to ensure safety and protection against the virus and fatal outcomes. Of course, not only the vaccinations are to blame for the different development across countries, but they are a crucial factor that users are very aware of and thus an easy one to blame. Furthermore, the sentiment towards AstraZeneca is even worse than that towards Pfizer in the UK. This development could be traced back to incidents of fatalities that have been connected to the vaccine.

However, it is important to note, that our interpretations should be taken with a grain of salt, as the limited amount of data collected has heavily impacted our findings.

# Legal and Ethical Perspective

When dealing with data, it is easy to forget that there are actually human beings behind the data and that it is our duty as data scientist to ensure that their personal data is handled with care, used with good intentions and to prevent any kind of abuse that may occur. When dealing with big data, there is no way we will not find ourselves in situations where we must inform ourselves about the legal scope of the project we are working on.

## Ethical Questions

Stoimenova, Rasheva-Yordanova and Christozov have identified volume, velocity and variety not only as defining factors for Big Data but also as factors which potentially allow for legal and ethical issues to arise. In their paper they describe how the gigantic volume makes it easier to obtain valuable bits of information that may need protection, how velocity allows for real time analysis and thus allows for immediate refining of a user's profile while the variety in the data can make users traceable. (Toleva-Stoimenova, Rasheva-Yordanova, & Christozov, 2018)

These were the main ethical questions whose solution we would like to explore now.

<u>How can we ensure we do not obtain sensitive or valuable data in our streaming process?</u>

This was one of the easier questions for us to answer over the course of the project. Since we are using tweets that are written by users themselves, they are in full control of what they post. Responsible

users are not highly likely to share information on social media because they are not comfortable with everyone having access to it. Thus, we argue that our streaming process does not catch data that could be used against the respective Twitter user. Furthermore, thanks to the fact that we only use the tweets for sentiment analysis, any sensitive data will be obscured by the classification as positive, negative, or neutral and does not have a chance to come to the surface within the scope of our project.

How do we obtain data that does not allow for refinement of a user's profile nor tracing?

We circumvented these two problems as well with an intuitive approach, when streaming the data, we only extracted the information that was valuable for our analysis such as the location and the sentiment. We refrained from collecting any other information such as the username and processed the obtained data such that explicitly stated locations were transformed into the respective countries which allowed us to conduct our analysis but also made it impossible to trace back a certain tweet to a singular person. Furthermore, the refinement of a user's profile was not of interest for us, as our focal point was on grasping the big picture and not on the individual.

Further Ethical Issues: Addressing the Impact of our Project - Biases

Emmanuel Derman describes in his "Hippocratic Oath of Modelling" that it is important to acknowledge the potential impact of a certain project. Since the topic we are dealing with is rather delicate and we recognize that our analysis may have been influenced by factors out of our control, which makes it important to also have a look at the biases that may be present in our work. (Kirrane, 2021)

First, our analysis is based on the sentiment of Twitter users, which already limits the data collection to a relatively privileged part of the world's population, as not everyone has access to the internet. Furthermore, when looking at the world-wide picture we could only capture English-speaking tweets which may exclude people lacking language education in non-English-speaking countries.

Another principal issue to address is that Twitter users are not required to be truthful, they may have not been vaccinated or they might not describe their experience with COVID-19 vaccinations in an inaccurate way. An example for this could be someone trying to influence others by tweeting about vaccines without having been vaccinated.

More aspects that influence the big picture from our analysis are technical issues such as the fact that the code for the sentiment analysis is troubled with sarcasm, the fact that we only streamed data for 5 days or that the algorithm sometimes includes tweets which, according to the words specified in the tracker, would fit the category in question but are actually about something completely different – an example for this would be that we wanted to include Data about the Johnson & Johnson vaccine and ended up having a tweet about the actress Dakota Johnson in our dataset.

## Legal Questions

Over the course of our project, we encountered several legal questions which required our attention. As tweets are the centre of our analysis, we will first outline the legal issues we encountered there, how we overcame them and how we ensured to comply with all the requirements needed to use the Twitter API. Afterwards we will address the handling of licenses, how we tackled citation and referencing and what steps we took to comply with EU guidelines.

Which aspects of The Twitter Developer Agreement apply to us and are we able to follow the required steps to ensure compliance with the statement we agreed to?

As the Twitter Developer Agreement mostly concerns developers of apps and services that use Twitter data, a lot of the concerns do not directly affect us in the same way. Nevertheless, we examined it closely and whenever we were not 100% sure that it did not apply to us, we made sure to comply with the requirements. Some of the most pressing questions we asked ourselves will be outlined below.

First, it is crucial to address the legal situation we find ourselves in when handling data streamed via the Twitter API. When applying for the developer access, we agreed to the "Twitter Developer Agreement". The version we followed is effective as of March 10, 2020. The license granted to us via accepting the Developer Agreement, is a *non-exclusive, royalty free, non-transferable, non-sublicensable, revocable license.* This allowed us to use the Twitter API to conduct analysis of the content posted on the platform, copy and display data obtained via the API, and modify content for display purposes (Twitter Inc, 2020). The use of the content is tied to certain conditions that were central to our project. Thus, we made it a priority to ensure every requirement is complied with.

To analyse the sentiment of tweets in different countries, we needed to collect the location from where the Tweet was sent. Section E. Location Data of the Agreement allowed us to clear some of our concerns:

> ***E. Location Data****. You will not (and you will not allow others to) aggregate, cache, or store location data and other geographic information contained in the Twitter Content, except in conjunction with the Twitter Content to which it is attached. You may only use such location data and geographic information to identify the location tagged by the Twitter Content. You may not use location data or geographic information on a standalone basis. (Twitter Inc, 2020)*

As we used the location data to group the streamed tweets by countries, we only extracted the location from Twitter such that it was still attached to the tweet itself. In the following analysis, the geographic information was reduced to a country as we omitted the specific location. When the sentiment scores were extracted from the tweets and when we conducted our further analysis, the location information was always connected to a component of the tweets content, such as the sentiment. Geographic information is also a part of privacy and public data which we will come to discuss again shortly.

Furthermore, we realized that for Twitter, the security of their API is of uttermost concern. Section G deals with this problem in more detail and we have included the first part of it below:

> ***G. Security****. You will maintain the security of the Twitter API and will not make available to a third party, any token, key, password or other login credentials to the Twitter API. (Twitter Inc, 2020)*

We made sure to follow the security requirements meticulously and did not share our Tokens/Passwords with anyone. When communicating them among our group we used end-to-end-encrypted messengers which allows for secure exchange by only granting our group access to the all the keys etc. that are communicated.

Another section we paid attention to deals with consent and permission by the users. We realized that in the scope of our project, the described actions were not applicable as they mostly deal with taking actions on the users' behalf which we abstained from and thus were not directly confronted with. However – we were directly affected by the off-Twitter matching issue. This refers to the fact, that a third person should not be able to match the data we obtained to any natural person based on the information stored in the data. We are thus required to either obtain consent to these sensible pieces of information or to not include any personal information apart from what Twitter provides us with.

In this category falls “public data” which includes e.g., profile information and display name as well as publicly stated location which is what we use for our analysis and thus do not interfere with the previously stated limitation on the use of location data.

As we are concerned with the privacy and protection of individual Twitter users, we also made sure that we comply with the Surveillance, privacy, and user protection section of the agreement. As the scope of our project is limited to sentiment analysis and a comparison of vaccination numbers and the public response they cause, we concluded that we do not infringe the stated restricted activities such as credit/insurance risk analysis, investigating or tracking sensitive groups and organizations, such as unions or activist groups, background checks or any form of extreme vetting or facial recognition ( (Twitter Inc, 2020)).

How do we ensure that additional resources we used – specifically the dataset for vaccine data - are handled according to the applying licenses.

Apart from the Twitter data, we used a public dataset that is updated daily and informs us about Coronavirus (COVID-19) vaccinations per country. We found this data on the website humdata.org. By reading their FAQ section we found out more about their Data Licenses and that everyone sharing data on their website shall use one of the licenses listed on their website (https://data.humdata.org/about/license). The applying license in our case is the “Creative Commons Attribution for Intergovernmental Organisations (CC BY-IGO)” which grants us the freedom to share, copy, transform the dataset in any way we wish as long as we give appropriate credit and provide source and author while not applying any legal terms that would hinder others to do the same. This type of license allows us to do everything we initially intended to. We only made simple transformations on the data, as we merely performed an extraction of those observations that were useful for our project. However, finding the original author and the appropriate referencing proved to be challenging which led us to email the site to find answers. We received a detailed explanation which, again, proved that the way we used the dataset was appropriate and the email also included the preferred way of citation which is the following (as also included in the reference section afterwards):

Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021). https://doi.org/10.1038/s41562-021-01122-8

Additionally, we used images or graphics from the internet for the design of e.g., the title page or the architecture graphic where we paid attention that they are free for personal use

EU Data protection

Since we are currently studying at a university located in a Member State of the EU, we also want to have a look at the legal situation in the European Union. Specifically, we were looking at the REGULATION 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL which entails the Directive 95/46/EC – also known at the GDPR, the General Data Protection Regulation. What we found out is that it is a fundamental right for a natural person to have their personal data protected. The GDPR is extensive and covers all kinds of situations, what we want to focus on is, whether – according to the EU – we are allowed to use the data twitter provides us with.

First, we were looking at situations where our project could be affected – for example §32 of the directive deals with consent of the concerned party. Since our data comes from twitter users, we need to find out whether those users consent to their tweets being processed as data. As indicated on twitters help page, when signing up, the user consents to the use of their personal data for the purpose

of "*Developer products, including our APIs and embeds*" which is the category our project fall into – thus we argue that we are within the scope of the EU's requirement in this regard (). Additionally, Twitter also has its own regulations in place to ensure their compliance with the GDPR where they focus on their product, policy and transparency (Twitter Inc, 2018).

Another interesting paragraph is §35 which deals with medical data, it could be argued that our project collects the vaccination data of some twitter users but when looking at §26 all our concerns are cleared up as it states:

> *The principles of data protection should* ***therefore not apply to anonymous information****, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that* ***the data subject is not or no longer identifiable****. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes. (Directive 95/46/EC, 2016)*

This means that since our data is anonymized in the streaming process, we do not face the risks of infringing the directive set out by the EU.

Citation and Referencing

As we were using external sources for our project, it is essential (not only in the field of data science but in all of academia) to properly cite and reference any sources. In terms of using code from other developers or websites such as stackoverflow, we included a link to the respective site in the Jupyter notebook to indicate that the respective chunk of code is not our intellectual property and/or that we took inspiration from the users referred to. Since the notebooks do not allow for a proper bibliography, an additional section in this Final Report PDF part of the final project serves as reference list for the programming part of the project. Unless requested otherwise by the creator, citations in the report part of the project, were created following the APA format as is common for our field of study.

## Discussion on the challenges encountered

To begin with, our team encountered minor challenges mainly associated with the architecture used for the project. As all the team members were new to Spark/Kafka, we had some difficulty setting up the streaming structure and making sure it works accurately. Initially, we have spent much time working out the errors and adjusting the framework, based on the code provided during the lectures and labs. While firstly testing various options and code snippets, some of the team members faced some issues with their access to Jupyter, which was caused by the lack of memory, fortunately we received a quick response from support team, who immediately provided us with more space. Besides, we faced some problems while saving the tweets, as the output files were initially messy and untidy, but the issue was addressed by calibrating some of the parameters and adjusting the code.

The first main coding issue, that we experienced, was a part of our idea to implement pre-processing and natural language processing algorithms directly into streaming procedures within PySpark framework. However, our code did not work initially, and we were constantly getting various errors, therefore we tried to implement the code from scientific articles found on the internet. The problem was easily solved during the lab session, where we were provided with a perfectly working piece of code, which we used as a base for our notebook. We also had to implement some more pre-processing, which turned out to be relatively complicated. As, for example, transforming the date column using PySpark did not work out initially, therefore we had to spend some time testing various code options to gather a tidy version of data.

In addition, due to the specificity of our project, we experienced some issues while deciding on how to store and analyse our data. Initially, we considered only live operations on data, but then we decided to accumulate the tweets to try various machine learning algorithms more conveniently. Due to this, we were searching for efficient ways to combine CSV with tweets into aggregated files and had to produce a small script for that. In our Jupyter environment, the individual CSV-files can be accessed via dp2-2021s-teamboehm/Test/TestCombine/combined_*vaccine-name_region*. The combined CSV files were handed in with the project.

As our project proposal implied deeper research into vaccination-related data within various countries. Therefore, we had to pick the countries with high enough vaccination rates (to get enough data for sentiment) and to consider the corresponding types of vaccines. However, the streaming process turned out to be more complicated, as Twitter users, located in various places (US, UK, Worldwide) were not active enough from time to time. This can be justified by some countries using only one type of vaccine or users being active only in some specific time windows. To deal with that, we decided to stream in various time windows, to define the most productive way to stream the data. The streaming procedure was also troublesome as we were often getting biased data tables, the folders with data had to be cleaned and new data had to be collected.

Despite the lack of activity of the users mentioned previously, we did not expect such a small number of tweets about Johnson & Johnson and Moderna vaccines, therefore we were forced to use only worldwide data for the analysis on these types of vaccines, which led to insignificant project limitations in terms of general analysis

We also encountered some difficulty with missing data, which was not streamed during one of the planned days due to some misunderstanding within the team, therefore we had to apply some programmatic methods to replace missing values. (In our case, an interpolation technique was used to accurately fill in the data for the average polarity of Moderna tweets).

To gain a deeper insight into the researched sentiment (also for our visualization), we decided to apply NLTK mechanisms, which turned out to be new for our team. We were challenged by some of the transformations and operations that had to be adjusted to our sentiment for more accurate results. For example, some of the tweets contained many useless symbols and words even after processing stages, so we had to play around with the parameters, stop words and functions to find the mistakes. Moreover, some functions from new libraries were to be implemented, therefore we had to research into their functionality and the hyperparameters, which had to be adjusted correspondingly.

We were also concerned with the number of streamed days, considered for the machine learning algorithms. Obviously, due to time constraints, we could not work with data, collected for months, therefore the linear regression model applied is not significant enough. Nonetheless, the model was set up correctly, so provided some more data is used, one can produce highly meaningful results.

Overall, the challenges encountered required some advanced and detailed research into the topic but in the end those challenges were the ones that contributed the most to our understanding of the employed tools. They forced us to look behind the façade and be able to understand the processes behind the different applications which might have been tedious at first but also rewarding in the end.

# Experience gained

## Katharina Böhm

Looking back, what I really like about this project is that it seems to be the first meaningful and more serious university project I have done so far. After having completed it, I feel proud and would show it to other people.  Even though the actual results of the project are faulty.

However, I must confess that I have not always had so positive sentiments about this project; classification: negative. Jokes aside, when discussing our project proposal as a team, we had many ideas about what we could do, but as none of us had worked with Kafka and Spark before, we were not sure what we would be able to do and what not. Considering the lacking knowledge, I was very doubtful and insecure in the beginning whether we would even get close to doing a project as proposed. However, as we started working on the project, things got clearer and clearer, especially with the help of my team members and stack overflow. I gained a better understanding of the different processes (streaming, cleaning, visualization, etc.) and the two packages. However, I am very much aware that there is still a lot to learn about Kafka and Spark and indeed feel extremely motivated to do so.

Nonetheless, my biggest take-away from this project are not necessarily all the technical details I have learned, but rather the experience of taking on a project without having a clue about the tools first. It is re-assuring to see, what one can do when being confronted with the seemingly impossible.

Furthermore, I am rather proud to say that the final version of our project still resembles our project proposal and that we did not have to alter it too much.

## Eva Jüngling

Starting off this project, I have to admit I was feeling a bit hopeless. My only experience in programming was the previous Data Science courses that had already challenged me quite a bit and this course was no different. While the content of the lecture and the labs was super interesting, I often had troubles following along due to the (indeed) fast-paced nature of the class. When writing our project proposal, we had a lot of different ideas, but we were still somewhat unsure of what our learning outcomes from the classes would be, so it was kind of a shot in the dark, which contributed to this feeling. Now that we have finished the project, however I can proudly say that we have exceeded my expectations.

The whole process of generating our own data out of “thin air” (aka twitter) via the magic of spark and Kafka was for some reason mind blowing to me and really made me hungry for more because it opened up this whole new world of big data that I haven’t been in contact with yet. The streaming process made me realize how powerful these tools are and suddenly our project seemed actually feasible. Having the experience of at first feeling like something cannot be achieved to realizing that we are actually capable of doing it, was a really great and reassuring one and motivated me for what’s to come in the future.

Furthermore, the legal and ethical side has been of great interest to me due to my 10 ECTS specialization in European and International Economic Law. I had previously not given the topic much thought in combination with data science but now that I had time to confront myself with the implications it is undeniable that they are tightly connected. Exploring this new perspective and applying it in the scope of our project really made the project feel like “something real” that one could also pursue outside of the classroom. It actually raised my interest so much, that I ended up stumbling

across a podcast called the “Data Science Ethics Podcast” which has allowed me to find a way to include data science in my daily life and broaden my horizons in that regard.

The last experience I gained through this project was, that I really appreciated the dynamic in the group – my team-members were so motivated, competent and reliable in a way that I have never experienced in a groupwork at university before. Previous groupworks in other subjects have often left me wishing I could have just done it on my own because the people I was working with were so unmotivated. It was completely different here - this experience has showed me the perks of groupwork and how efficient it can be if everyone is able to focus on what they do best and actually try to achieve good results which is something I am very grateful for.

## Hashym Ulukhanov

Having started the assignment, I was extremely new to the concepts and tools that are required to deal with big data projects. Moreover, the project milestones, that our team highlighted during some planning sessions, looked totally infeasible at first. I did not have experience with Spark/Kafka before, therefore right after the first lecture, I was intensively reading some articles and numerous discussions to gain some deeper insight into the topic. Nevertheless, the next lectures and labs turned out to be extremely informative and helpful, as I was basically getting the code and the theory, that I was could accurately apply to get the output.

In my personal opinion, the streaming process was a bit more complicated to set up at first, however I enjoyed the big data feature of the project, therefore I spent much time playing around with spark functions to become more aware of Spark’s functionality. I also have to admit that the spark architecture is quite complicated to understand, and the documentation provided was not helpful enough, therefore I had to read some articles about PySpark. Apart from the Spark framework, I became more confident with Machine Learning algorithms and NLP, especially after applying some of the tools in the context of project.

Generally, I encountered various challenges during the project, which mainly include the errors associated with new libraries and functions used, but they definitely helped me to get acquainted with many new methods and techniques (interpolation for missing values, for example).

All in all, the aforementioned issues effectively provided me with an intensive introduction to big data science, while also improved my overall programming skills. I acknowledge that such group assignments offer a great opportunity to experience the real working environment by introducing one to the inner structure of data science projects.

## Peter Tavaszi:

In this part of the report, I am supposed to write about the challenges that I encountered personally during this project. However, first, I want to set forth that I enjoyed this project from the beginning to the end because I personally like difficult challenges like this one. It was quite unusual and unexpected for me how much we as students had to figure out on our own for this subject.

I had to acknowledge the fact that I did not have so much previous knowledge about the topics discussed in this course. I also had to accept that we as a group had to write a research assignment about complex topics and that I was going to have to invest a lot of time into the project. On the one hand, I had to realize that a lot of additional background research is expected in this subject and that

we have to figure out a lot of new stuff and develop many new skills on our own. On the other hand, I realized that this teaching method gave us very much freedom in terms of how we want to approach things like sentiment analysis, streaming tweets from Twitter, creating our own data, visualizing our results, etc. And to be honest I liked that we could experiment on our own as much as possible.

I was responsible for setting up the Twitter streaming process, building the sentiment analysis pipeline, performing the sentiment classification models and the linear regression models for observing the effect of average sentiment polarity on daily vaccinations, creating the word clouds, and writing the description of the project solution part of this report. I had to solve which felt like a million different issues and error messages on my own and every time I overcame one of these annoying problems, it felt like a small accomplishment or a little new skill that I acquired. These experiences made me stronger every time I had to face other problems, which I have never ever seen before, and they clearly made me more resistant to giving up and letting someone else fix these issues. I generally like to have some basic previous knowledge or a rough plan before plunging into a large project or facing a tough challenge like this.

I benefitted from this course exceptionally because I gained an entirely new understanding of how it is to create something entirely from scratch and having to experiment with new tools that one has never thought of using one day. I once again proved to myself that with strong determination and dedicated work everything is possible. I always set the bar high at everything I do and even if this project did not turn out exactly as I expected, I am very happy that I had the opportunity to be part of this experience and that I learned so many valuable skills. I really liked that the evaluation for this course was built on practical implementations and a hand-in assignment, because I think it is way better than having to study for a final exam and then forgetting everything.

I am very proud of our real time big data processing project, and I am looking forward to similar new challenges and the possibility to fine-tune my newly acquired skill set. Lastly, I am a bit disappointed that we did not learn more about the foundations of sentiment analysis in this course and that we did not touch upon the topic of fraud detection which seems to be an interesting practical implementation of real time processing and machine-learning algorithms.

## Recommendations for future work

The biggest issue we could identify was the limited number of days and tweets we received when streaming. To achieve better results and to be able to observe in which direction the sentiment is really going, the streaming process should probably run for multiple hours a day over the course of weeks while also taking the different time-zones into consideration.

Furthermore, some of the biases should be tackled. The pandemic is, as is already suggested by the name, an international problem. Limiting the scope of tweets to English-speaking users only, required us to remove a crucial part of the world's population that would have significantly aided in understanding the overall reception of the vaccine better. We used English to simplify the project, but other regions, such as India that are currently battling the delta variant should be considered as well, thus the Natural Language Processing algorithm should be extended to include more languages and to assess the sentiment in other regions of the world – at least in those where the data promises to be helpful or interesting. We have not yet found a solution to the user-based problem of untruthfulness, when continuing to work with tweets in the future, we must accept these issues. However, we have concluded, that even if the tweets are not based on real events, they still represent a real sentiment of the user and thus this "bias" is acceptable for us.

Perhaps the biggest issue that would need to be addressed in future work would be the classification as Positive/Negative/Neutral during the process of the sentiment analysis. When looking at the files, it was often apparent that (to the human eye) clearly negative/positive tweets have not been classified as such and were stated to be “neutral”. An approach to tackle this issue could be to manually inspect and classify the data which then could serve as training data to improve the sentiment analysis process. Certain abbreviations, sarcasm, slang, etc. were not detected and by including supervised learning mechanisms into the equation, better results could be achieved. This also ties in with the classification issue itself, the best scenario would be to limit it to two classes – positive and negative – to have the option to apply as many models as possible (such as the Logistic Regression which was not able to find any meaningful results when applied in this context).

Finally, it is crucial to consider the legal perspective of the project when creating further analysis following the same pattern. As we are enrolled in a university course and received access to the Twitter API based on agreeing that we are not showing this project to anyone except our teachers, the legal situation would be an entirely different one when the project is continued in a different setting. The implications and requirements of the Twitter Developer agreement and Privacy Policy must be revisited and implemented to ensure compliance and an ethical and legal process.

Once having collected this bigger dataset that is less biased and classified in a more meaningful way, we trust that the application of our models would lead to improved results that would help finding ways to better understand the public reception of the pandemic and the measures taken against it and maybe it would even help to improve the process of how vaccination is handled in the world. We see potential in our work and are curious as to whether the results could be better than what we managed to achieve if more data was employed and evaluated using the different models.

# References Report


Dean, B. (2021, February 10). (Backlinko LLC) Retrieved July 2021, from backlinko.com: https://backlinko.com/twitter-users

HDX. (2021, July 2021). Retrieved July 2021, from data.humdata.org: https://data.humdata.org/organization/hdx

O'Dowd, A. (2021, June 21). Covid-19: Cases of delta variant rise by 79%, but rate of growth slows. *British Medical Journal Publishing Group, 373*.

Rosa-Aquino, P. (2021, July 2). *https://nymag.com/intelligencer/tags/covid-19/*. Retrieved July 2021, from https://nymag.com: https://nymag.com/intelligencer/2021/07/will-the-j-and-j-vaccine-need-a-boost-against-delta.html

Toleva-Stoimenova, S., Rasheva-Yordanova, K., & Christozov, D. (2018, November). NEW DIMENSIONS OF DATA SCIENCE PROFESSIONAL SKILLS AS EMERGED BY IDENTIFIED ETHICAL ISSUES: GDPR. *11th annual International Conference of Education, Research and Innovation*, (pp. 488-497). Sevilla.

Twitter Inc. (2020, March 20). *Developer Agreement and Policy – Twitter Developers*. Retrieved July 2021, from Twitter Developers: https://developer.twitter.com/en/developer-terms/agreement-and-policy

Twitter Inc. (2020, June 18). *Twitter Privacy Policy.* Retrieved July 2020, from Privacy Policy: https://twitter.com/en/privacy

Twitter Inc. (2021). *Twitter's GDPR Hub*. Retrieved July 2021, from twitter.com: https://gdpr.twitter.com/en.html

Kirrane, S. (2021). *Lecture 1*. Data Processing 2: Scalable Data Processing, Legal & Ethical Foundations of Data Science. Vienna University of Economics and Business.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (2016) *Official Journal* L119, p. 1.

# References Graphics and Images Report

**Title page:**

Image. By Wallpaper Cave, 2021. Retrieved from: https://wallpapercave.com/analytics-wallpapers

**Architecture Graphic (these have been altered to achieve the desired appearance**

Apache Kafka Logo. By nicepng, 2021. Retrieved from: https://www.nicepng.com/ourpic/u2w7t4r5y3o0o0u2_kafka-logo-tall-apache-kafka-logo/

CSV Logo. By Awawers, 2021. Retrieved from: https://awawers.net/assets/uploads/csv.png

Classification Clipart. By Noun Project, 2020. Retrieved from: https://thenounproject.com/term/classification/

Jupyter Logo. By Project Jupyter, 2021. Retrieved from: https://jupyter.org/assets/main-logo.svg

Sentiment Analysis Clipart. By Eva Jüngling, 2021.

TextBlob Logo. By Steven Loria, 2020. Retrieved from: https://textblob.readthedocs.io/en/dev/#

Twitter Logo. By Stick PNG, 2021. Retrieved from: http://www.stickpng.com/img/icons-logos-emojis/tech-companies/twitter-logo

Visualization Clipart. By vectorified, 2021. Retrieved from: https://vectorified.com/command-line-icon

**Screenshots and Visualizations:**

Created by ourselves using Python

# References Jupyter Notebooks

## Covid Vaccination Dataset:

Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021). https://doi.org/10.1038/s41562-021-01122-8

## Additional Resources for Coding

Classification and regression - Spark 3.1.2 Documentation. (2021). Apache Spark. https://spark.apache.org/docs/latest/ml-classification-regression.html

Extracting, transforming and selecting features - Spark 3.1.2 Documentation. (2021). Apache Spark. https://spark.apache.org/docs/latest/ml-features.html

Li, S. (2018a, May 1). Building A Linear Regression with PySpark and MLlib. Towards Data Science. https://towardsdatascience.com/building-a-linear-regression-with-pyspark-and-mllib-d065c3ba246a

Li, S. (2018, July 20). Multi-Class Text Classification with PySpark. Towards Data Science. https://towardsdatascience.com/multi-class-text-classification-with-pyspark-7d78d022ed35

Prabhakaran, S. (2018, October 2). Lemmatization Approaches with Examples in Python. Machine Learning +.

".S.". (2019, December 19). Sentiment analysis on streaming twitter data using Spark Structured Streaming & Python. Github. https://github.com/stamatelou/twitter_sentiment_analysis

Vu, D. (2019, November 8). Generating WordClouds in Python. DataCamp. https://www.datacamp.com/community/tutorials/wordcloud-python

## Stackoverflow:

Burt, (2016, July 12 13:08). Best way to get the max value in a Spark dataframe column. Stackoverflow. https://stackoverflow.com/questions/33224740/best-way-to-get-the-max-value-in-a-spark-dataframe-column

Jochen Ritzel, (2010, August 29 13:09). Finding the common elements of a list. Stackoverflow. https://stackoverflow.com/questions/3594740/finding-the-common-elements-of-a-list

Paul, (2019, September 21 9:24). Removing punctuation in spark dataframe. Stackoverflow. https://stackoverflow.com/questions/58038919/removing-punctuation-in-spark-dataframe

Saleem Khan, (2020, January 23 2:34). nltk wordnet lemmatization with POS tag on pyspark dataframe. Stackoverflow. https://stackoverflow.com/questions/59850159/nltk-wordnet-lemmatization-with-pos-tag-on-pyspark-dataframe

Werner, (2021, July 3 21:09). Pyspark: non-matching values in date time column. Stackoverflow. https://stackoverflow.com/questions/68239001/pyspark-non-matching-values-in-date-time-column