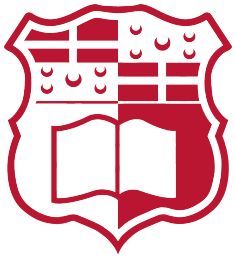


# European Covid-19 Vaccine Sentiment Analyses On Twitter and Article Data


**Aiden Williams**

Supervised by Dr Claudia Borg

Department of Artificial Intelligence

Faculty of ICT

University of Malta


**June, 2021**

*An Individual Assigned Practical Task submitted in partial fulfilment of the requirements for the degree of B.Sc. in Artificial Intelligence.*

# FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as “the unacknowledged use, as one's own, of work of another person, whether or not such work has been published, and as may be further elaborated in Faculty or University guidelines" (University Assessment Regulations, 2009, Regulation 39 (b)(i), University of Malta).

I / ~~We~~*, the undersigned, declare that the [assignment / Assigned Practical Task report / Final Year Project report] submitted is my / our* work, except where acknowledged and referenced.

I / ~~We~~* understand that the penalties for committing a breach of the regulations include loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

* Delete as appropriate.

(N. B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

| | |
|---|---|
| Aiden Williams |  |
| Student Name | Signature |
| | |
| Student Name | Signature |
| | |
| Student Name | Signature |
| | |
| Student Name | Signature |

| | |
|---|---|
| ARI2201 | European Covid-19 Vaccine Sentiment Analyses On Twitter and Article Data |
| Course Code | Title of work submitted |

# Contents

# List of Figures

# List of Abbreviations

# 1

# Introduction

## 1.1 | Abstract

Modern social media has entrenched itself as the most accessible and heard voice for the modern digitized populace. Platforms such as Twitter, Facebook and YouTube allow their users to generate and share opinionated content across the Web. For this reason, the scale of the content in this available data makes it difficult to access and use. However, less than 1% of Twitter data is geo-tagged, which means that classifying the origin country of any data gathered is tedious. For this purpose of geo-localization, newspaper articles are better suited, as it is easier to identify for which country an article is written for. Progress made in the Natural Language Processing (NLP) field has made data collection, analyses and evaluation on big data language projects possible. Sentiment Analyses (SA) has become a popular tool used for automated analyses for large scale and varied language content. By implementing a model that is able to give a sentiment score, it would become possible to analyze and extract knowledge from raw data collected from social media platforms and news article headings.

## 1.2 | Aims and Objectives

Positive public opinion on the Covid-19 vaccine is essential for an effective vaccine roll out and public or private strategies that are affected by the pandemic. The aim of this Individual Assigned Practical Task (IAPT) is to collect publicly available opinions on the Covid-19 pandemic, analyze and extract valuable knowledge to present as an explanation to the sentiment change over time. The effect of Twitter data and publicly available newspaper article headings was investigated for trends and insights.

A number of objectives were defined for this IAPT:

1. Collect a sizeable and relevant publicly available dataset from a social media platform that covers a number of European Union (EU) member states.
2. Collect a sizeable and relevant publicly available dataset from a number of European newspapers that covers a number of European countries.
3. Preprocess and translate this data so that it can be used as an input for a SA model.
4. Implement a SA model to analyze the collected data.
5. Analyze and extract knowledge from the SA scores by visualizing the results.

## 1.3 | Summary of Solution Developed

Using the Twitter Application Programming Interface (API) and by manually using the Google search engine to collect a number of articles related to six EU countries: the United Kingdom, Spain, France, Germany, Italy and the Netherlands. A number of Python notebooks were developed to collect, filter, process, analyze and visualize this data and the products of its analyses. The Google Cloud Translate API was used to translate non-English tweet texts to English texts. A Valence Aware Dictionary and Sentiment Reasoner (VADER) (Hutto and Gilbert, 2014) model was used as provided by the Natural Language Toolkit (NLTK) library (Bird et al., 2009). Finally, the multiplex visualization library (Mamo, 2021) was extensively used to visualize the data in its processed and analyzed form.

# 2

# Background & Literature Overview

## 2.1 | Background

The Covid-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic first identified in December 2019 in Wuhan, China. As of 12 June 2021, more than 175 million cases have been confirmed, with more than 3.78 million confirmed deaths attributed to COVID-19, making it one of the deadliest pandemics in history. Late in 2020 vaccination programs started to roll out in various countries. In January 2021, countries that form part of the EU started their vaccine programs. In December 2020 the European Commission published a survey titled: Public Opinion on Covid-19 Vaccination in the EU 2020 (eup, 2020). In this survey, a total of 24,424 interviews where conducted in over 27 EU countries to see the public's views on:

1. Attitudes to vaccination.
2. General satisfaction with EU measures to fight the pandemic.
3. Information on Covid-19.

The data collected was done manually via either online forms or telephone calls and was represented visually in the published survey.

### 2.1.1 | Data Collection

For a proper analyses of online EU content, a varied dataset is collected. Two general sets of sources, social media platforms and online newspapers are identified as sufficient and adequate for this IAPT.

#### 2.1.1.1 | Social Media Platform Data Collection

Social media platform data collection can be done via several methods. Most platforms offer official Representational State Transfer (REST) API, while for others a traditional web-scraping techniques are used to collect the data. In 2016 Twitter users posted an average of 500 million tweets, daily (Crannell et al., 2016). In the same year, the number of active Twitter users exceeded 22% of the internet users in the world (Kayser and Bierwisch, 2016). This amounted to 342 million daily active users at the time. More recently this figure has changed to 330 million daily users (Tankovska, 2021).

Another popular social media platform is Facebook. In the case of Malta 92.3% of the population (nap, 2021) is an active user of the platform. The possibility of using Facebook as a data source was explored where Covid-19 related influencers, social leaders and local media pages' posts would be collected as well as their audience user engagement, .i.e comments and shares. However the API was found to not be as accessible or usable as the Twitter API. Due to its global reach and usable API it was decided that Twitter would be the social media platform from which public opinionated data will be collected. A large scale dataset of tweets is maintained by Banda et al. (2021).

#### 2.1.1.2 | Newspaper Article Data Collection

Unlike social media platforms the more traditional newspapers, sometimes referred to as the fourth estate, are decentralised and not limited to one platform. In fact many modern newspapers utilize social media to share and advertise on. During the research done for this IAPT it was found that most published works that use an article headline dataset, have this dataset collected manually such as in Chaudhary and Paulose (2019). However a number of online APIs where found and tested with the intention of collecting a dataset spanning 6 months. These APIs were newsapi (new, 2021a)) and newscatcher (new, 2021b), notwithstanding their topic and country search functionality, accessing articles past one month was a paid feature. So it was decided to not use these APIs due to budgeting reasons. It was then decided that the article heading dataset would be collected and validated manually.

### 2.1.2 | Data Processing

Europe is a diverse continent full of different peoples, cultures and languages. To match data with an origin country two methods where used. Either the data came with a geotag, .i.e the origin country was specified somewhere with the text of the data. Or the source language was used. Due to a number of different languages collected the deci-

sion had to be taken whether to implement a multi-lingual SA model or use machine translation. Balahur and Turchi (2014) state that commercial engines are able to translate from and into many languages. Araújo et al. (2020) use and experiment number commercially available machine translation engines. From the translators reviews the Google Cloud Translation was the cheapest and most applicable for this IAPT.

### 2.1.3 | Sentiment Analyses

As an active research field that has emerged recently, SA is a discipline that extracts people's feelings, opinions, thoughts and behaviors from user's text data using NLP methods (Danneman and Heimann, 2014). Moreover, SA is also known as opinion mining, with emphasis on text classification problem. Extracting sentiment information from web-scale text data can be very challenging and expensive task due to large amount of data (Fernández-Gavilanes et al., 2016). A survey conducted by Lo et al. (2016) explores two main approaches for SA, subjectivity and polarity detection. There subjectivity detection described as an understanding on whether the content contains personal views and opinions as opposed to factual information. Polarity detection as the study of subjectivity with different polarities, intensities or rankings. In the context of this IAPT polarity detection makes more sense than subjectivity detection. A frequently used and readily accessible SA model is VADER (Hutto and Gilbert, 2014).

# 3

# Materials & Methods

## 3.1 | Overview

For this IAPT a multi-stage approach was used for data collection, analyses and evaluation. This approach is described in figure 3.1 as a flowchart.

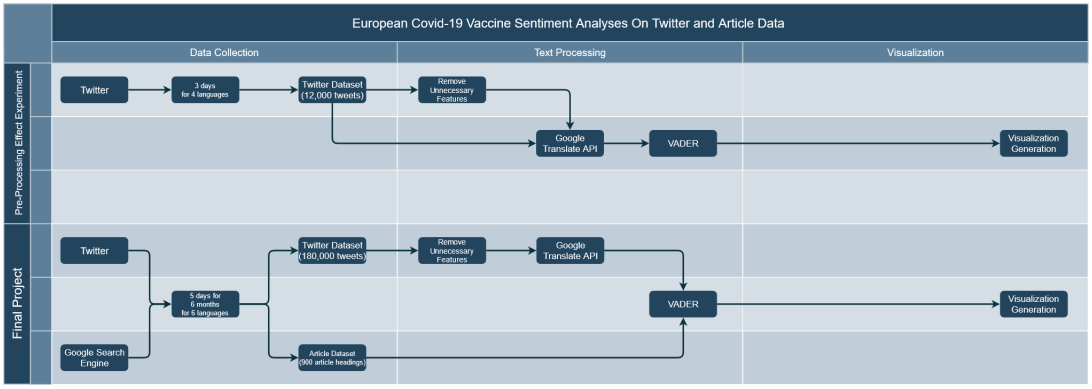



Figure 3.1: Flowchart of the Method of Experiment

## 3.2 | Data Collection

### 3.2.1 | Twitter Dataset

The large scale dataset of tweets maintained by Banda et al. (2021) is very large and analysing it in its entirety would be expensive, lengthy and out of scope for this IAPT, since the majority of the data present is generated outside the EU borders. The dataset contains daily folders representing each day from the 22nd of March 2020, which contain

two files, one of all tweets and retweets on the day, and the other a cleaned version with no retweets. Instead, all of this available data for every month from December 2020 till May 2021, 5 dates, the 1st, 7th, 14th, 21st, 28th where selected, and the clean dataset file was used. It was decided that 1000 tweets from 6 languages would suffice to form the Twitter dataset. The 6 languages, English, Spanish, French, German, Italian and Dutch are the European languages with the most presence in the Panacea Lab dataset. An important point to consider when using the Panacea Lab dataset is that only the tweet ID is available. However, this ID can be easily used to download and collect the tweet text and other tweet features. For this IAPT the Python library, Tweepy a Python wrapper for the Twitter API was used (Roesslein, 2020). The use of the Twitter API requires the signing-up to the service with Twitter, where after approval API are shared. The free version comes with limitations such as a limit of 300 lookups/ 15 minutes, however a limit on status retrieval through the ID does not exist. Hence, for this part of the data collection no limitations were imposed by the Twitter API. In total 180,000 tweets where collected. Figure 3.2 shows the distribution of the languages within a day from the full Twitter dataset.

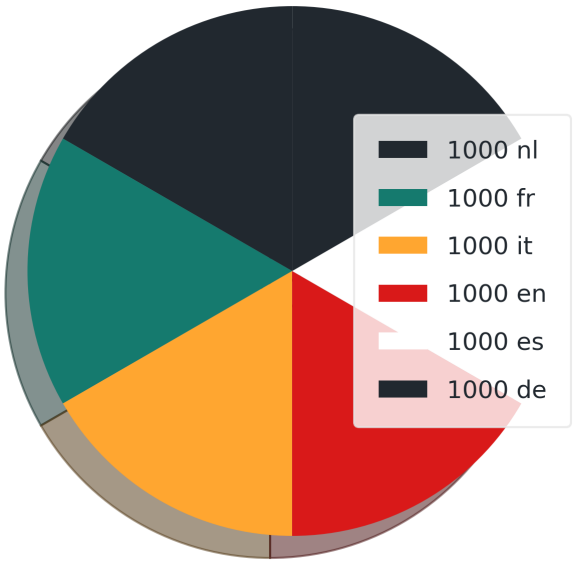



Figure 3.2: The language distribution in the Full Twitter Dataset

A smaller dataset was also collected following the same method as described above. This smaller dataset features 1000 tweets from 3 dates, the 1st of January, February and March and 4 languages, English, Spanish, French, German. This dataset was used to test the effect of pre-processing the tweets before passing them to a VADER model. In

total 12,000 tweets where collected. Figure 3.3 shows the distribution of the languages within a day from the experimental Twitter dataset.

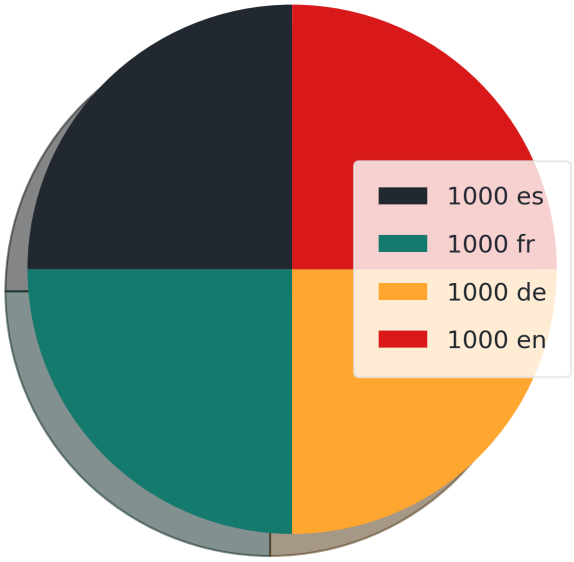



Figure 3.3: The language distribution in the Experiment Twitter Dataset

### 3.2.2 | Article Dataset

As the focus of this API was explained to be more directed towards using social media data, it was decided that the article dataset would be much smaller in scale. Using the Google search engine a query was done for every day a file was taken from the Panacea Lab dataset. The query was of the format:
{{Country} Covid* before:{Date in YYYY-MM-DD Format} After:{Day before Date in YYYY-MM-DD Format}}
5 relevant articles where collected for each day and stored in a number of csv files. In total 900 articles where collected.

## 3.3 | Data Processing

The tweet data collected is unfiltered and contain a number of useless features such as links, stop words and whitespace. In the first part of this IAPT the results of pre-processed tweets and 'naked' tweets are used to create a final pre-process function. In the final version tweets go through the following process:

- HTML special entities (e.g. &) are removed.
- Tickers are removed.
- Hyperlinks are removed.
- Punctuation is removed.
- "'s", "'t", "'ve" are split with a space.
- Words with 2 or fewer letters are removed.
- Whitespace is removed.
- Stop words are removed.

The stop word list for the various languages where taken from the NLTK (Bird et al., 2009) library. Article headings in the article dataset do not go through this pre-processing step.

For translation, non-English pre-processed tweet text was passed sent to the Google Cloud Translation API. The total cost of using the translation API came to €60. Article headings where not translated as they were collected only in English.

## 3.4 | Sentiment Analyses

The NLTK (Bird et al., 2009) library was employed to implement a VADER (Hutto and Gilbert, 2014) model. Each tweet and article heading was used as an input, and the compound SA score was kept. The mean compound score of each day was stored in a file, while each score was classifies as either:

- very positive (score >= 0.75)
- positive (0.25 <= score < 0.75)
- neutral (-0.25 <= score < 0.25)
- negative (-0.75 <= score < 0.25)
- very negative (score < -0.75)

## 3.5 | Visualization

The SA scores where visualized with the use of the Multiplex (Mamo, 2021) library, and the matplotlib library (Hunter, 2007) via the use of Time Series plots and word clouds for each month. A word cloud visualization was also generated for every language (translated to English) for every month. The focus of these visualizations was aimed at the results stemming from the Twitter dataset. An example of a time series plot generated can be seen in figure 3.4 while an example of a French word cloud for the month of May can be seen in 3.5. The complete collection of visualizations can be found in appendix A.

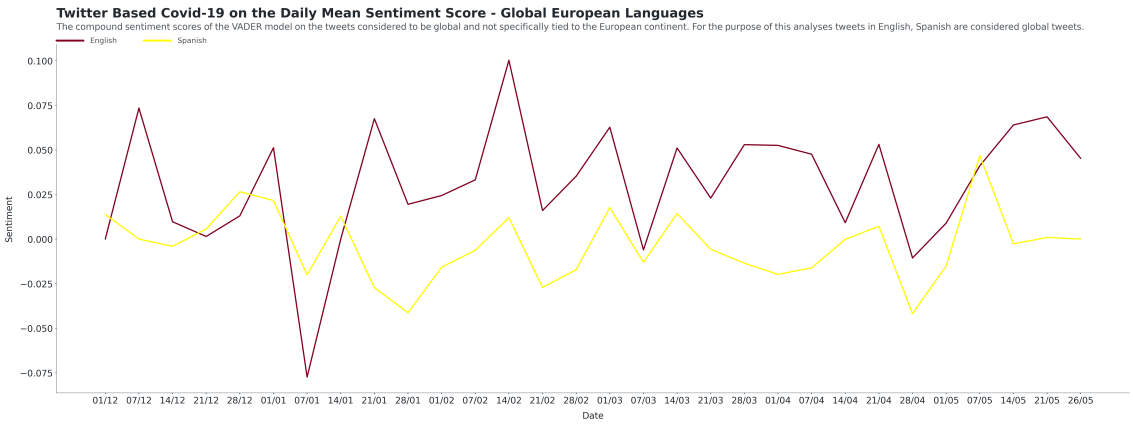



Figure 3.4:


Figure 3.5: French Word cloud for May

# 4

# Evaluation

The evaluation of this IAPT is based on a number of visualizations based on the output of the VADER model. In this evaluation the differences between the various visualizations are covered and interesting details and findings are explored. Since giving a detailed evaluation for each language would be very lengthy, this evaluation will focus on the French data while the rest of the visualizations can be found in Appendix A. The French data was chosen as it was found to be very close to other EU countries' sentiment such as Germany, Italy and the Netherlands, while tweets collected in English and Spanish where seen to be more closely related to global affairs than to European ones. Figure 4.1 shows the classification of the SA scores, plotted and annotated with article headings and Twitter tweets while figures 4.2 to 4.4 show the French word clouds for the months of January, February and April.

This form of visualization was found to give more information than by plotting the mean score of every day. As can be seen, peaks in a positive line usually correlates with a dip in the negative lines. By analyzing these peaks a sense of what was going on during that time period can be taken, especially when matched with a tweet or an article heading from that day. For example, on the 1st of January 2021, a peak in the positive lines can be seen, however when matched with a random tweet from that day, it becomes clear that spirits are high because of a new year, and not because the covid-19 situation is improving. Furthermore, for the rest of the month, the positive line can be seen going downwards, and the very negative line even has a peak on the 21st. When this is compared to the word cloud for January, a number of negative terms are shown such as 'disaster', 'fatal' and 'chronic', while the combination of the frequency of the terms 'proof' and 'vaccination' suggests that the French still want more proof for the effectiveness of the vaccine prior to approval. This idea is reflected in the survey conducted by the EU (eup, 2020) where the French are shown to be the least enthusiastic to

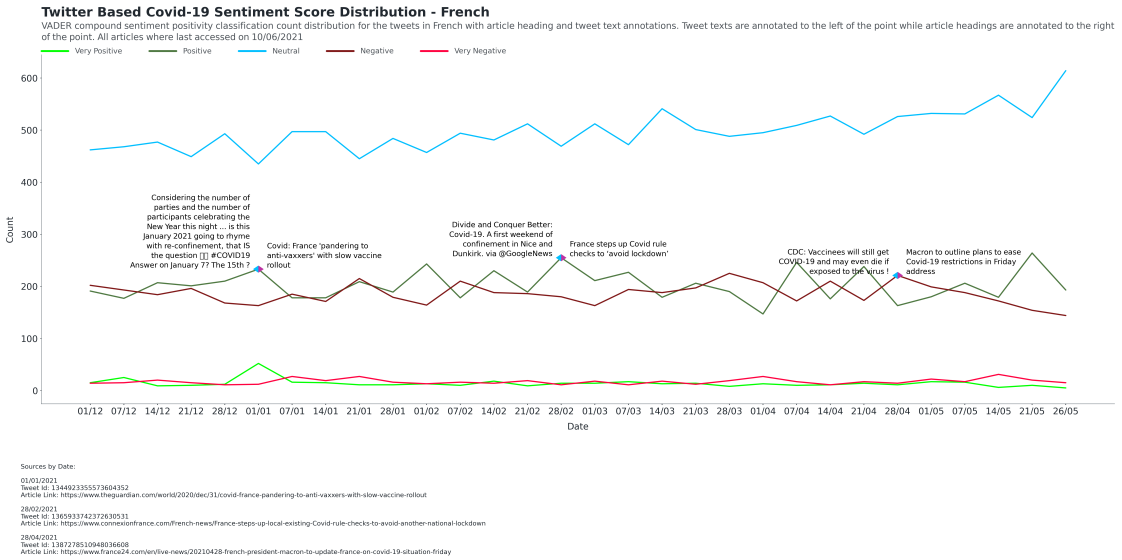



Figure 4.1:

get vaccinated. It is also an interesting date regarding vaccination as it marks the start for the EU program where most countries did not start vaccination before the start of 2021.

Another date, the 28th of February marks a point in time when France was experiencing a lockdown. The general sentiment of the time shows an increase in positivity, and the annotated text can be interpreted. When compared to the word cloud of February, the terms 'will' and 'proud' show an approval the situation.

After this point the count of neutral tweets start to increase, this can be interpreted as a sign that the Covid-19 content on Twitter is becoming calmer. Finally, on the 5th of April a peak in the negative line is observed. When analyzed, a number of tweets show expressed the lack of trust in the effectiveness of the vaccine. When compared to the word cloud of February, the terms 'forced', 'useless', 'ruins', 'vaccine' and 'miserable'. As the 5th marks a date of decreasing government restrictions it can be assumed that these terms indicate negative sentiment towards the vaccine.


Figure 4.2: French word cloud in Januray


Figure 4.3: French word cloud in February

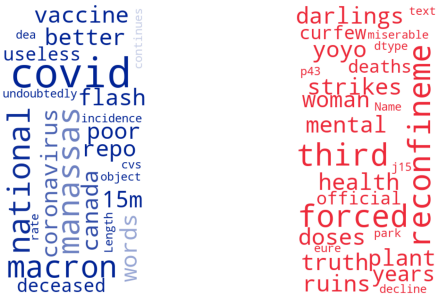



Figure 4.4: French word cloud in April

# 5

# Conclusions

## 5.1 | Achieved Aims and Objectives

Recalling the aim of this IAPT from the Introduction, it can be considered to be fulfilled as 2 datasets of European Covid-19 related data where collected in the form of a collection of tweets and articles. Analyses and extraction of valuable knowledge is also achieved in the evaluation and visually in Appendix A. Hence, the objectives of this IAPT can also be considered to be achieved as preprocessing was done with negligible loss of information from the raw data. The implementation of the VADER model was also satisfiable and several visualizations have been successfully produced.

## 5.2 | Critique and Limitations

Due to the amount of visualizations produced, evaluation was limited to only one language and was also not done fully. Some processes such as translation, and the use of some APIs such as the ones considered for collected the article dataset were not available for free. This is what lead to manually collecting the article dataset and not using an API.

## 5.3 | Future Work

In the future I would like to expand the analyses done by going over each language in detail observing how tweet and article data trail each other. Furthermore I would expand more datasets, particularly the article dataset.

## 5.4 | Final Remarks

This IAPT was a very interesting and fun assignment to work on. Achieving the objective of extracting knowledge from the data gathered was particularly challenging and time-consuming as the proper visualization technique took time to develop. However, by being able to extract trends and identify events from the data was satisfying.

# A

# Visualization



## A.1 | Data Collection Validation

In the pre-processing experiment, the method used for collecting and filtering the 1000 Tweets. To test that the distribution and collection was being done correctly, figures A.1 and A.2 where generated.

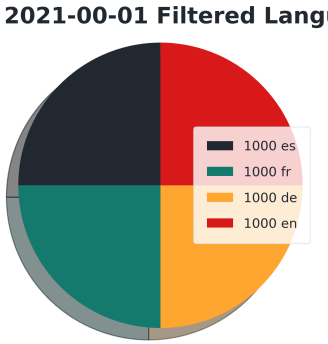



Figure A.1:

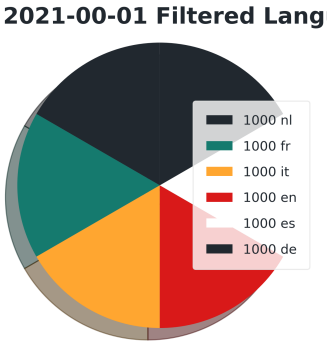



Figure A.2:

## A.2 | Pre-Processing Effect

The compound sentiment scores returned from the VADER model was averaged and plotted for the 3 dates used in the experiment. The 4 graphs produced, figures A.3 to A.6, show a time series graph of each language over the 3 days which are 1 month apart.

The difference calculated along with the shapes of the graphs plotted were not considered significant enough to remove pre-processing. However the pre-process function was amended to keep more features like emojis and hashtags, which prior to this experiment where being removed.

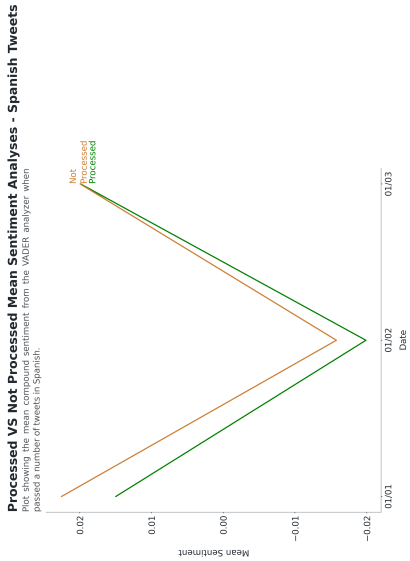



Figure A.3:




Figure A.4:




Figure A.5:




Figure A.6:

## A.3 | Daily Twitter Mean Sentiment

Figures A.7 to A.10 time series graphs where plotted for the mean sentiment scores on the larger 180,000 tweet dataset.

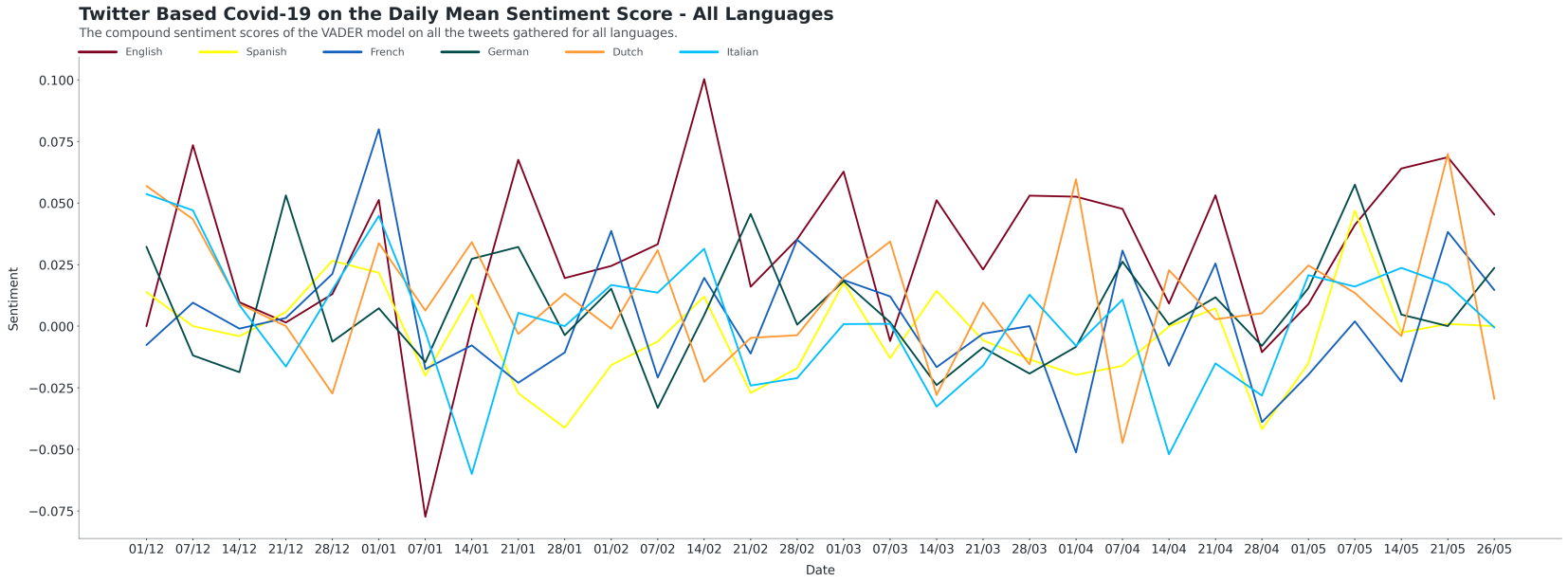



Figure A.7:

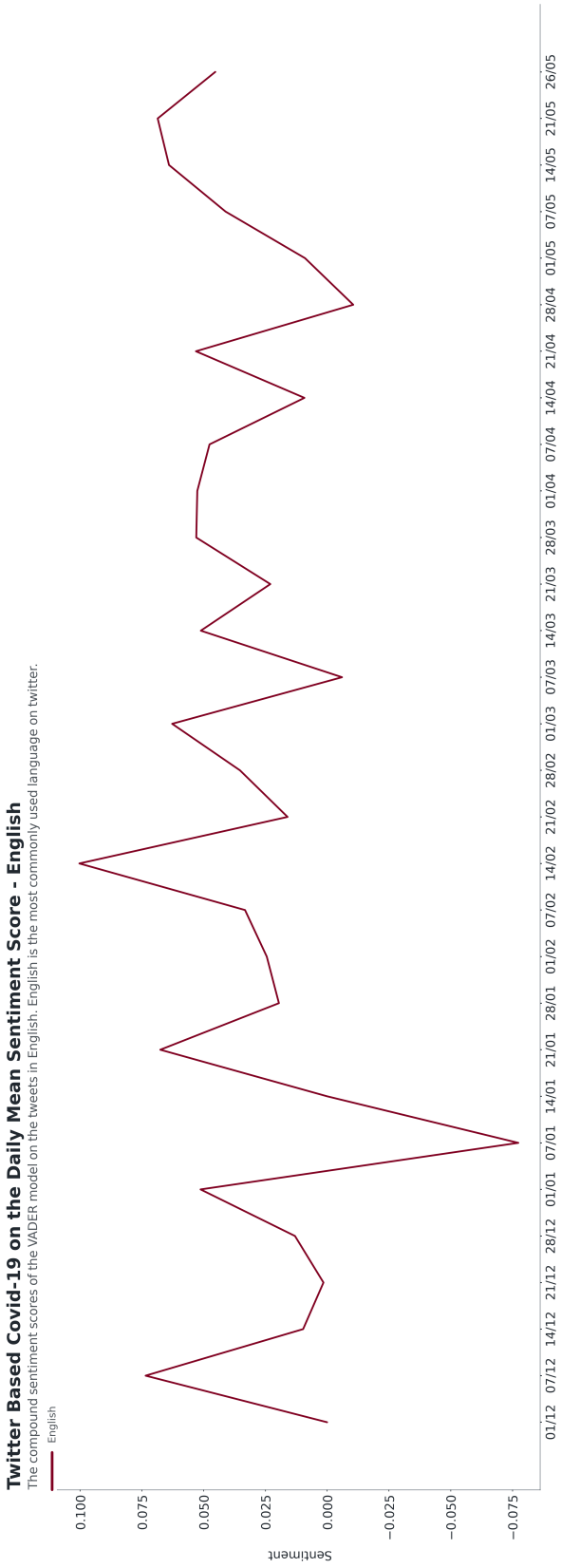

Twitter Based Covid-19 on the Daily Mean Sentiment Score - English
The compound sentiment scores of the VADER model on the tweets in English. English is the most commonly used language on twitter.
English
Sentiment
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05
Date

Figure A.8:

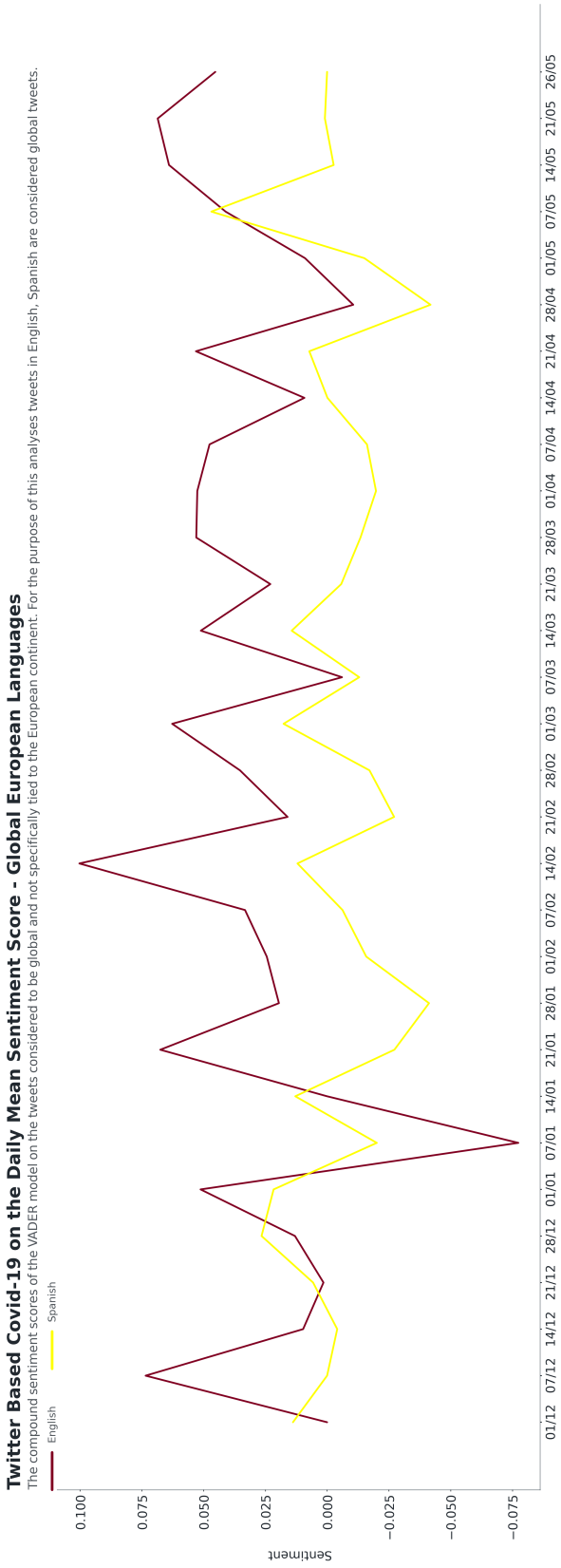

Twitter Based Covid-19 on the Daily Mean Sentiment Score - Global European Languages
The compound sentiment scores of the VADER model on the tweets considered to be global and not specifically tied to the European continent. For the purpose of this analyses tweets in English, Spanish are considered global tweets.
English
Spanish
Sentiment
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05
Date

Figure A.9:

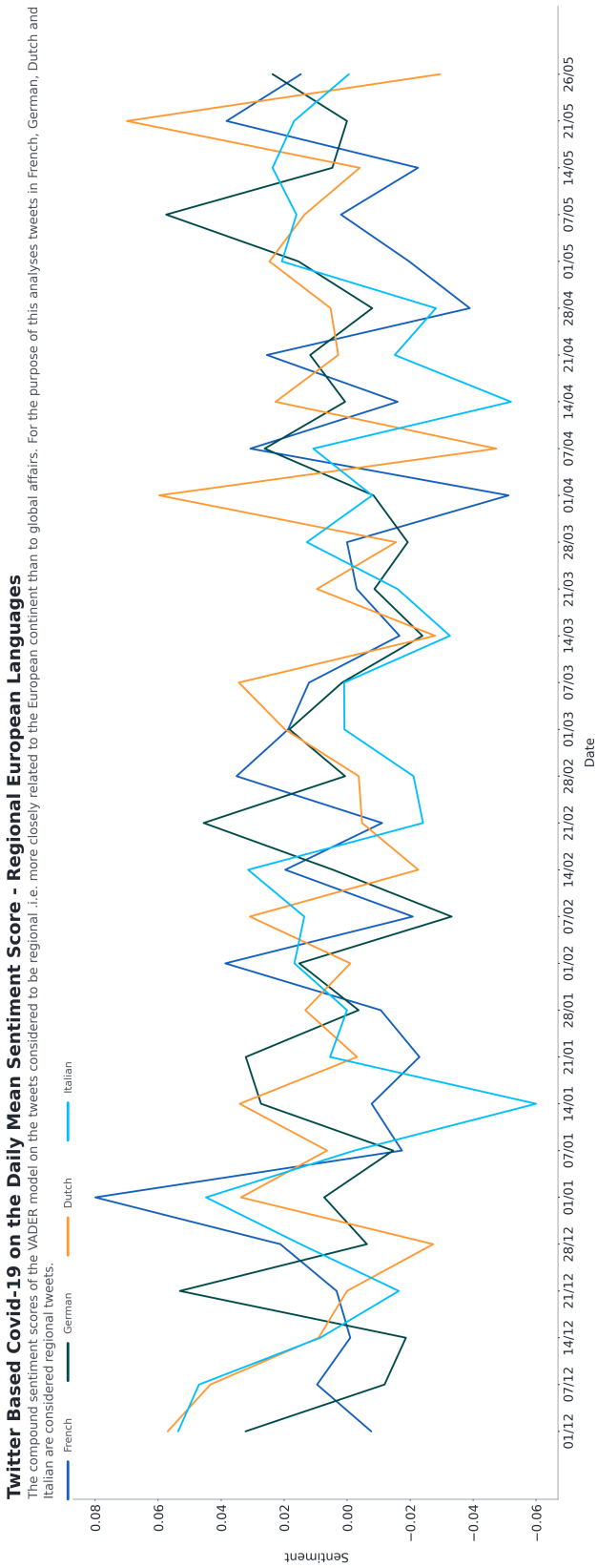
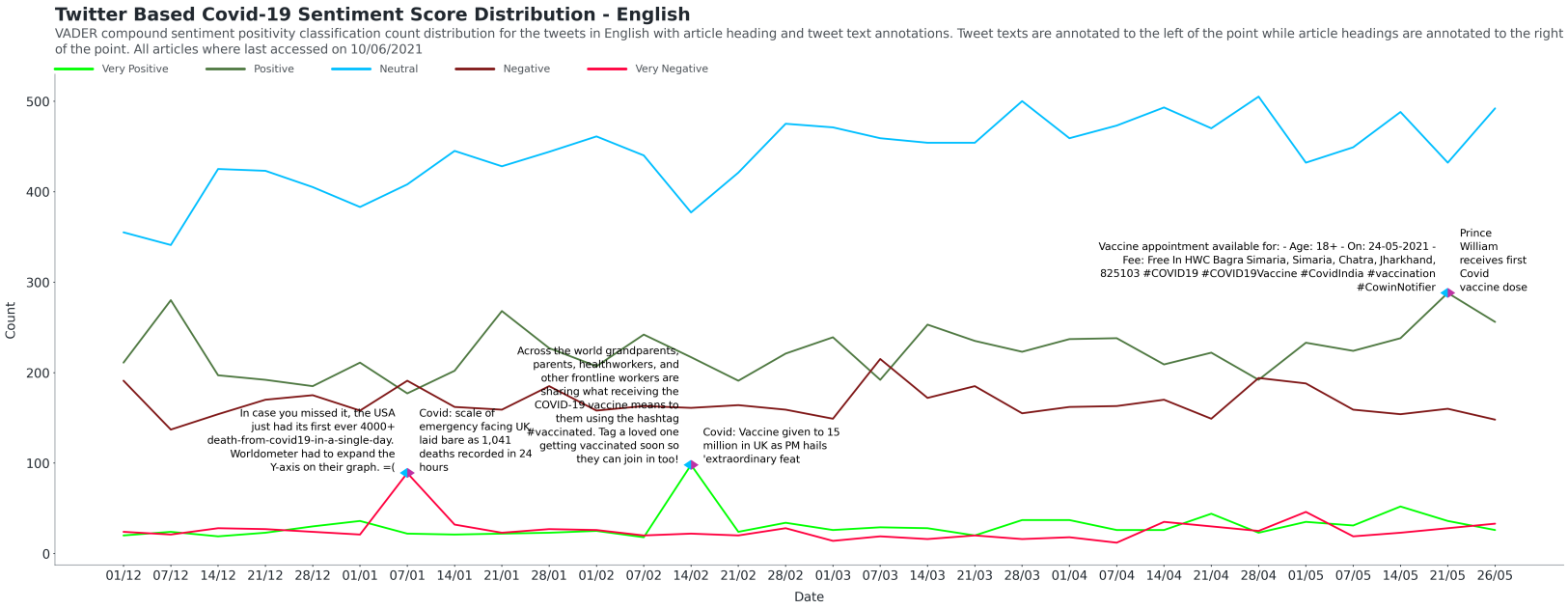

Twitter Based Covid-19 on the Daily Mean Sentiment Score - Regional European Languages
The compound sentiment scores of the VADER model on the tweets considered to be regional .i.e. more closely related to the European continent than to global affairs. For the purpose of this analyses tweets in French, German, Dutch and Italian are considered regional tweets.
French
German
Dutch
Italian
Sentiment
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05
Date

Figure A.10:

## A.4 | Daily Twitter Sentiment Classification

By classifying each compound score in a positivity class, a time series plot is plotted for each country can better visualize the changes sentiment over time. Article headings and tweets annotated to dips and peaks of the lines in the plot to represent a random snippet of what was being said on that day. Figures A.11 to A.16 show these plots.



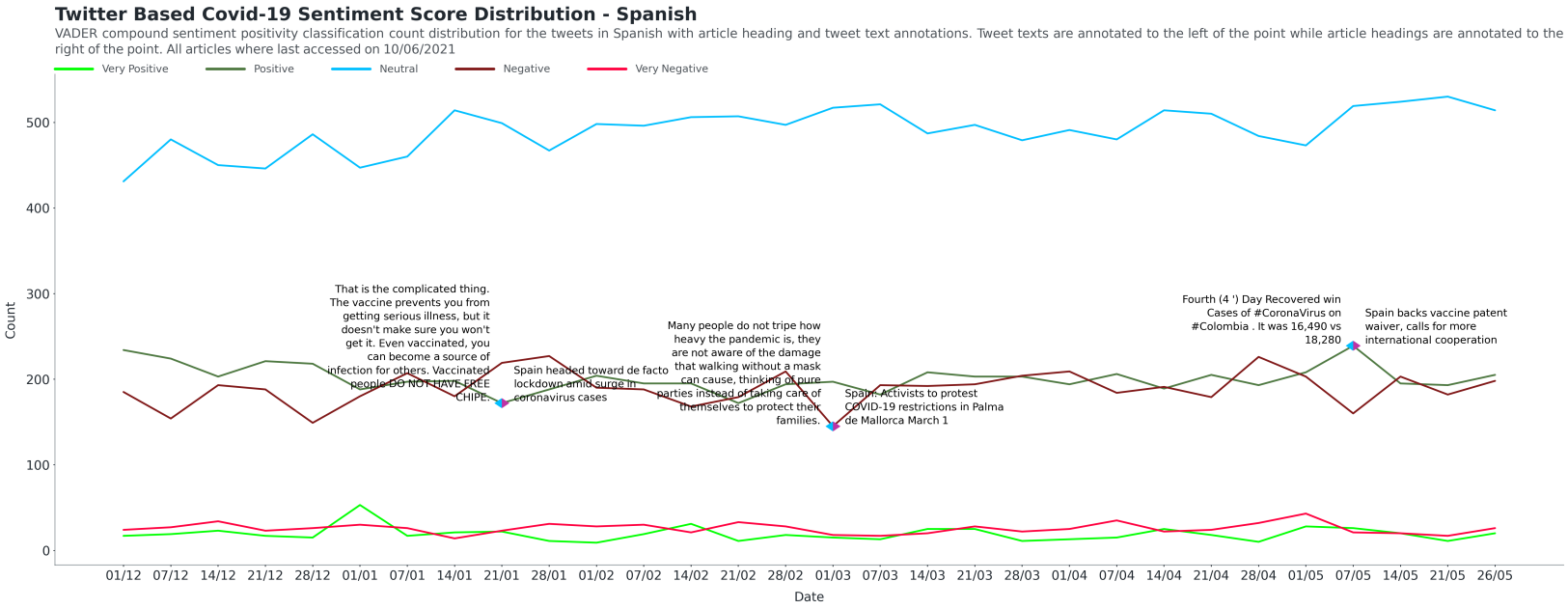

Sources by Date:

07/01/2021
Tweet Id: 1347045708461867009
Article Link: https://www.theguardian.com/world/2021/jan/06/scale-of-emergency-facing-uk-laid-bare-as-1000-die-in-24-hours

14/02/2021
Tweet Id: 1360819822112694272
Article Link: https://www.bbc.com/news/uk-56062976

21/05/2021
Tweet Id: 1393055120720482307
Article Link: https://www.bbc.com/news/uk-57185584

Figure A.11:



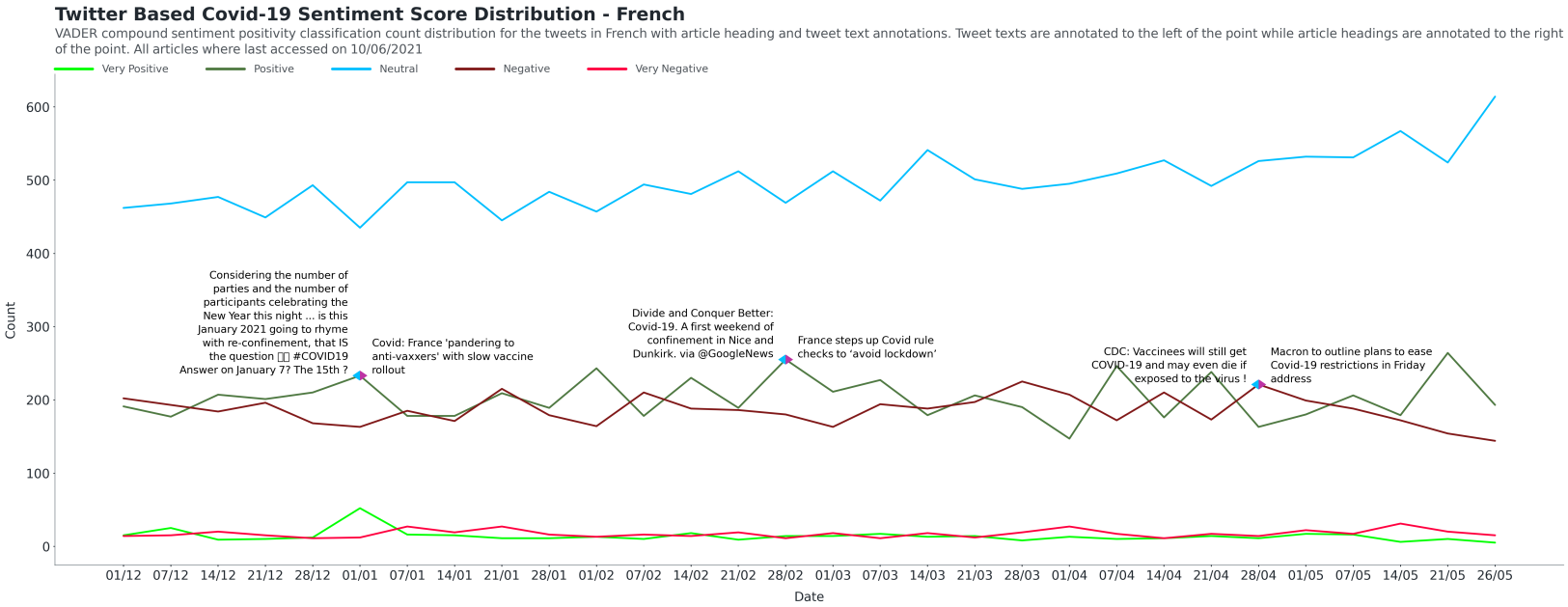

Sources by Date:

21/01/2021
Tweet Id: 1352122671614582784
Article Link: https://english.elpais.com/society/2021-01-20/spain-headed-toward-de-facto-lockdown-amid-surge-in-coronavirus-cases.html

01/03/2021
Tweet Id: 1366264385447088128
Article Link: https://www.garda.com/crisis24/news-alerts/449101/spain-activists-to-protest-covid-19-restrictions-in-palma-de-mallorca-march-1

07/05/2021
Tweet Id: 1390528805916262408
Article Link: https://www.reuters.com/world/europe/spain-says-vaccine-patent-waiver-is-way-forward-not-enough-poor-countries-2021-05-06/

Figure A.12:



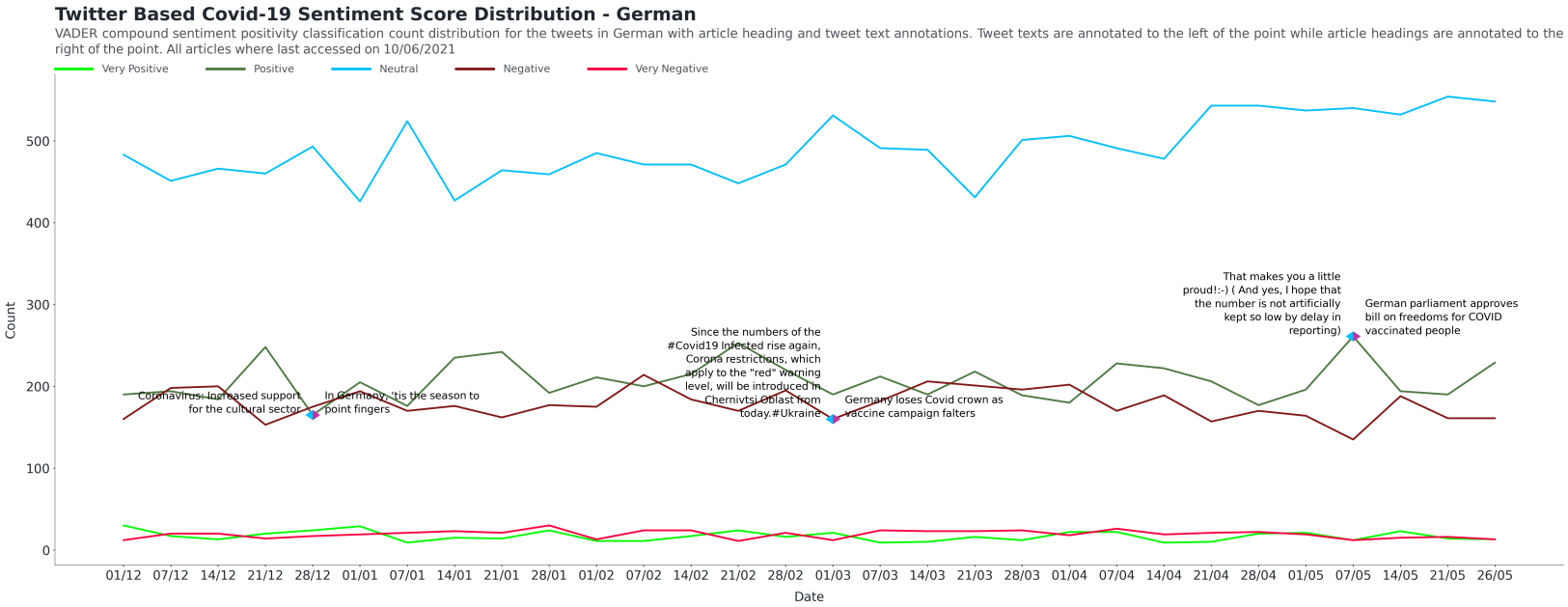

Sources by Date:

01/01/2021
Tweet Id: 1344923355573604352
Article Link: https://www.theguardian.com/world/2020/dec/31/covid-france-pandering-to-anti-vaxxers-with-slow-vaccine-rollout

28/02/2021
Tweet Id: 1365933742372630531
Article Link: https://www.connexionfrance.com/French-news/France-steps-up-local-existing-Covid-rule-checks-to-avoid-another-national-lockdown

28/04/2021
Tweet Id: 1387278510948036608
Article Link: https://www.france24.com/en/live-news/20210428-french-president-macron-to-update-france-on-covid-19-situation-friday

Figure A.13:



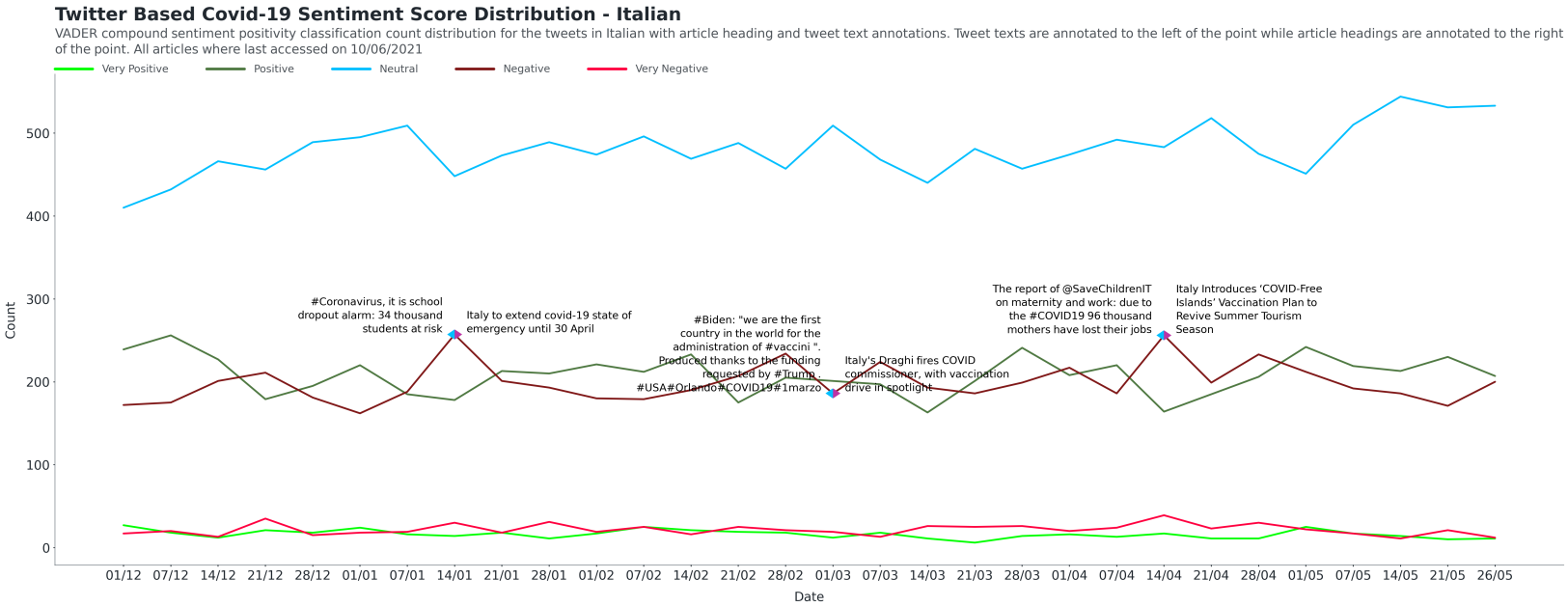

Sources by Date:

21/12/2020
Tweet Id: 1340926578860109832
Article Link: https://www.politico.eu/article/germany-angela-merkel-coronavirus-lockdown-restrictions-crisis/

01/03/2021
Tweet Id: 1366278288977649670
Article Link: https://www.ft.com/content/33f8ffd6-066b-449c-bf7e-edd51d661b19

07/05/2021
Tweet Id: 1393115883510419457
Article Link: https://www.dw.com/en/german-parliament-approves-bill-on-freedoms-for-covid-vaccinated-people/a-57450855

Figure A.14:




Sources by Date:

14/01/2021
Tweet Id: 1347081626468163584
Article Link: https://www.wantedinrome.com/news/italy-to-extend-covid-19-state-of-emergency-until-30-april.html

01/03/2021
Tweet Id: 1366308492122931200
Article Link: https://www.reuters.com/article/us-health-coronavirus-italy-commisioner-idUSKCN2AT2U9

14/04/2021
Tweet Id: 1390566764384243717
Article Link: https://www.schengenvisainfo.com/news/italy-introduces-covid-free-islands-vaccination-plan-to-revive-summer-tourism-season/

Figure A.15:

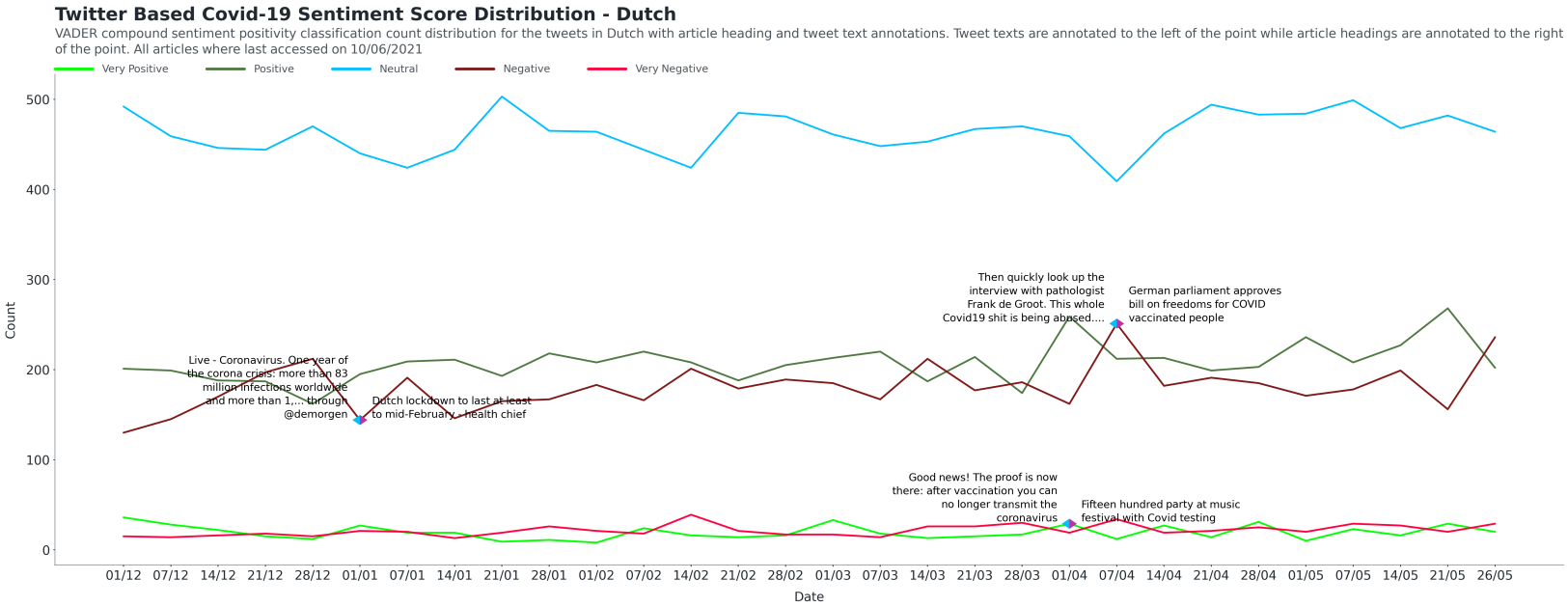



Figure A.16:

## A.5 | Daily Article vs Twitter Mean Sentiment

To see the relationship news articles and tweets have the daily mean sentiment scores of each language where plotted. Figures A.17 to A.22 show these plots.

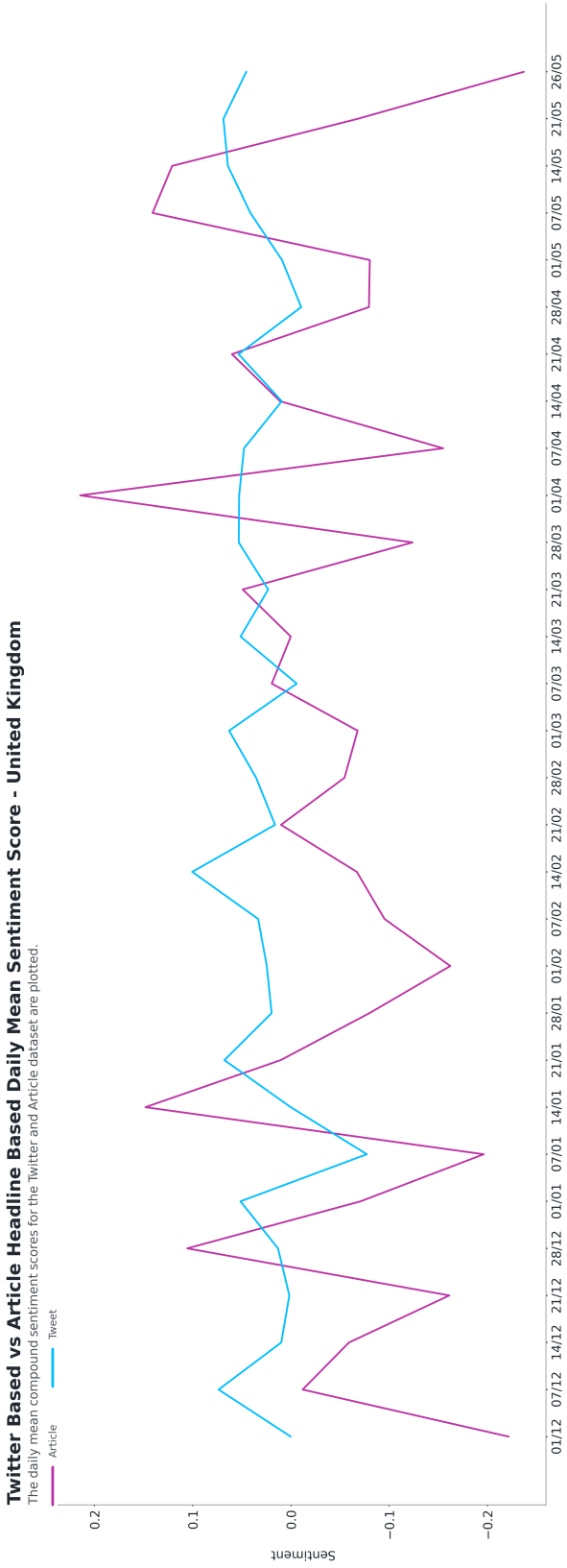

Twitter Based vs Article Headline Based Daily Mean Sentiment Score - United Kingdom
The daily mean compound sentiment scores for the Twitter and Article dataset are plotted.
Article
Tweet
Sentiment
0.2
0.1
0.0
−0.1
−0.2
Date
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05

Figure A.17:

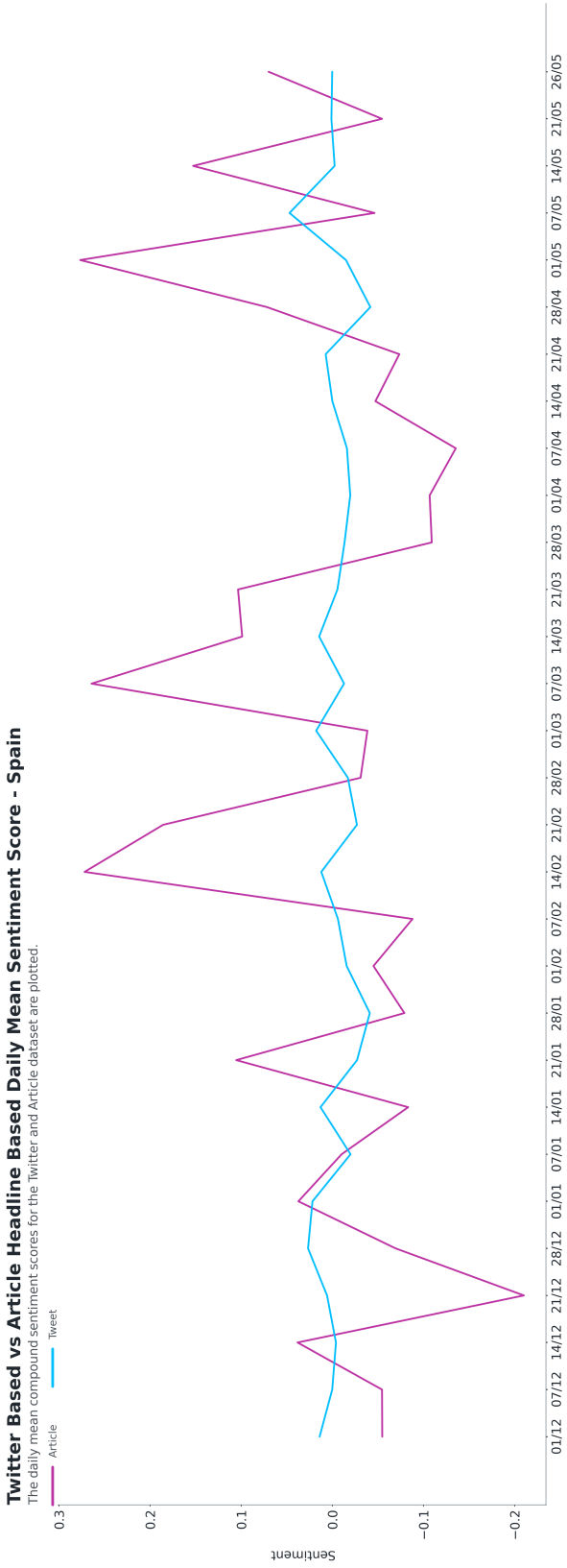

Twitter Based vs Article Headline Based Daily Mean Sentiment Score - Spain
The daily mean compound sentiment scores for the Twitter and Article dataset are plotted.
Article
Tweet
Sentiment
0.3
0.2
0.1
0.0
−0.1
−0.2
Date
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05

Figure A.18:

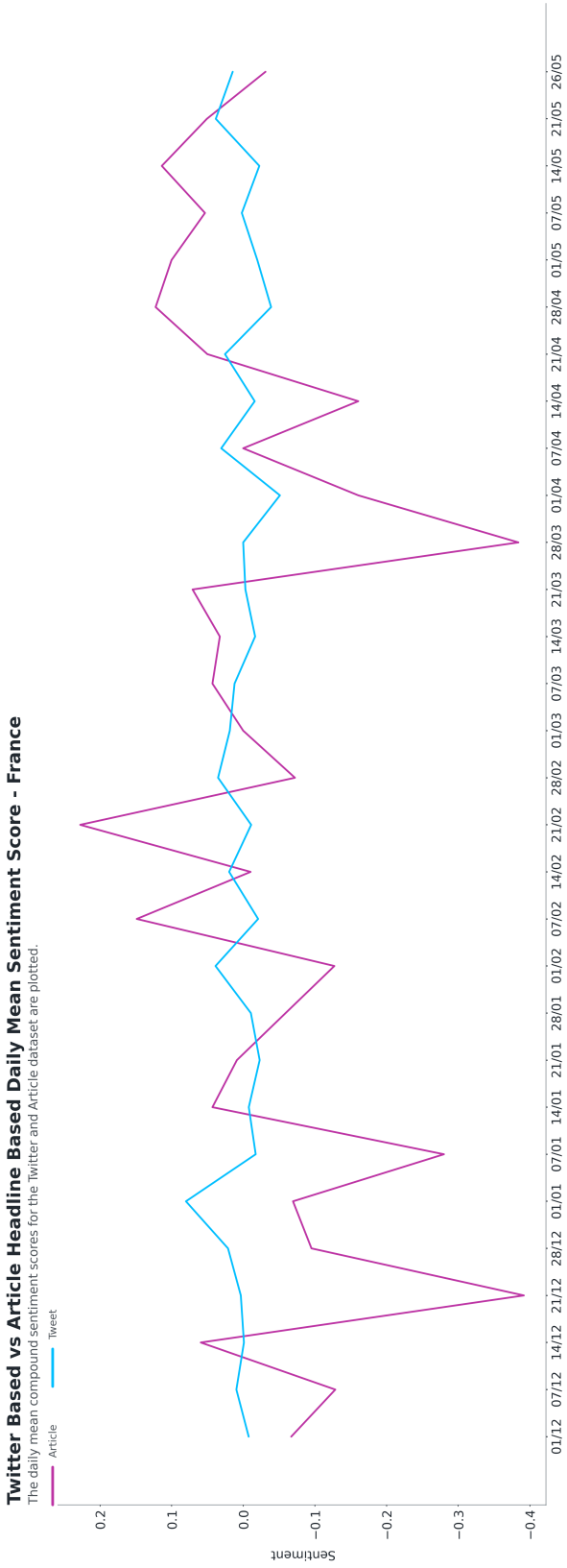

Twitter Based vs Article Headline Based Daily Mean Sentiment Score - France
The daily mean compound sentiment scores for the Twitter and Article dataset are plotted.
Article
Tweet
Sentiment
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05
Date

Figure A.19:

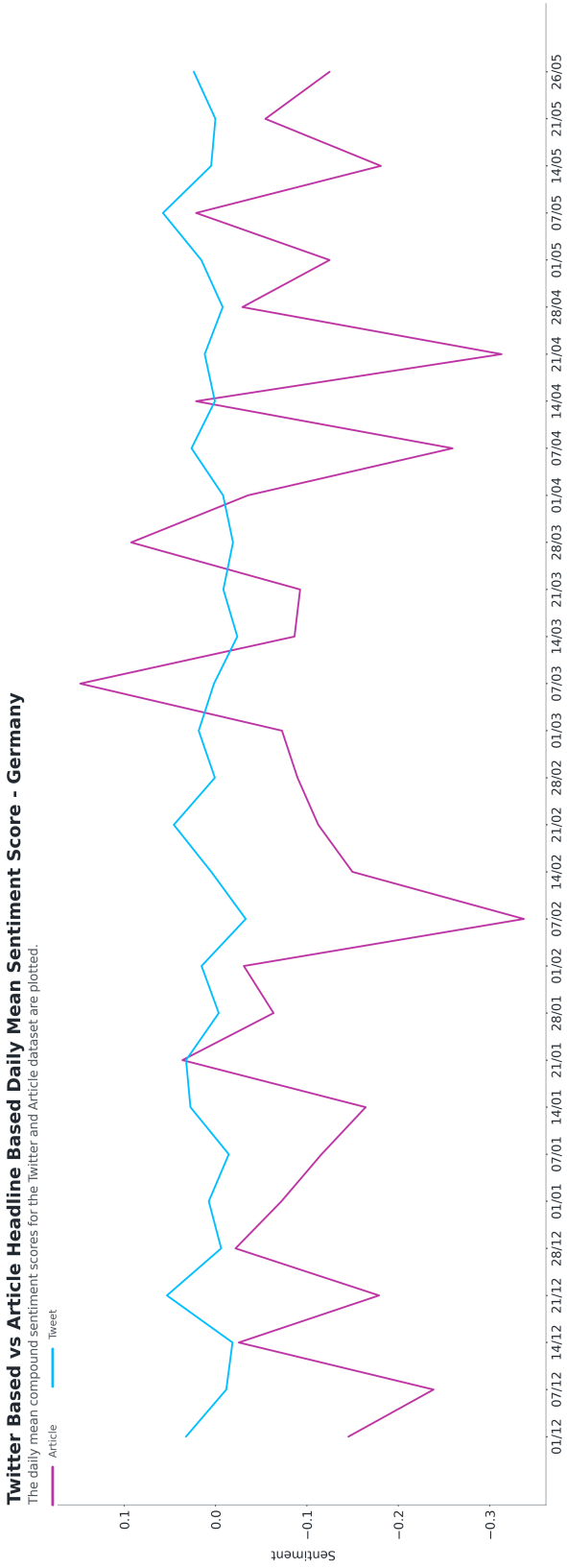

Twitter Based vs Article Headline Based Daily Mean Sentiment Score - Germany
The daily mean compound sentiment scores for the Twitter and Article dataset are plotted.
Article
Tweet
Sentiment
0.1
0.0
−0.1
−0.2
−0.3
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05
Date

Figure A.20:

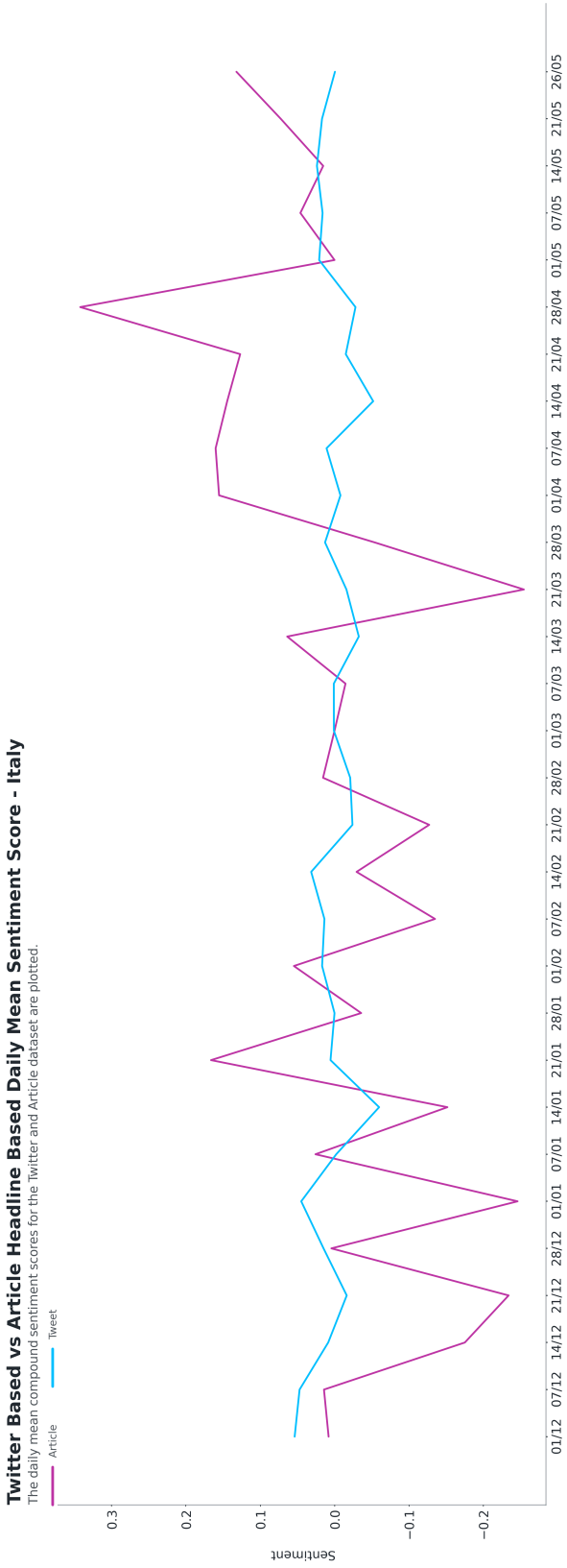

Twitter Based vs Article Headline Based Daily Mean Sentiment Score - Italy
The daily mean compound sentiment scores for the Twitter and Article dataset are plotted.
Article
Tweet
Sentiment
0.3
0.2
0.1
0.0
−0.1
−0.2
01/12
07/12
14/12
21/12
28/12
01/01
07/01
14/01
21/01
28/01
01/02
07/02
14/02
21/02
28/02
01/03
07/03
14/03
21/03
28/03
01/04
07/04
14/04
21/04
28/04
01/05
07/05
14/05
21/05
26/05
Date

Figure A.21:

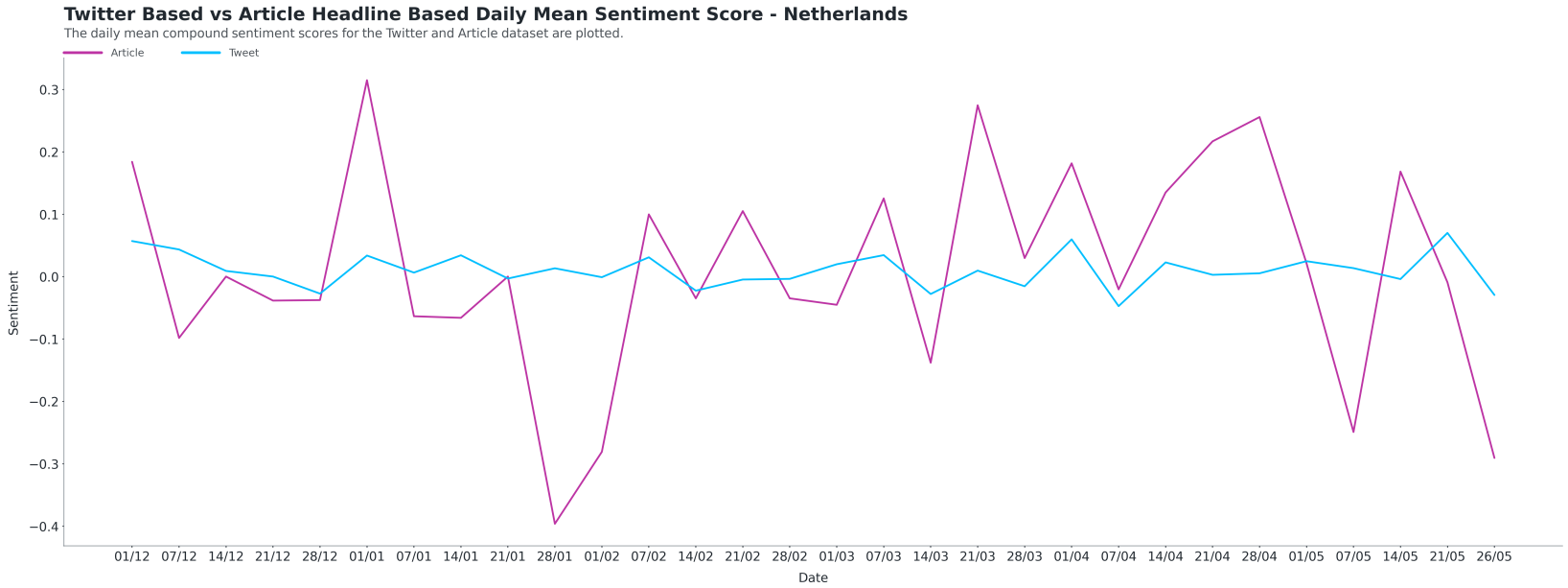



Figure A.22:

## A.6 | Word Frequency in the Form of Word Clouds

Finally a word cloud was generated for each month in each language. These word clouds where used to follow the difference of word usage over the 6 month period. The remaining figures A.23 to A.58 show these word clouds.


Figure A.23: English word cloud in December

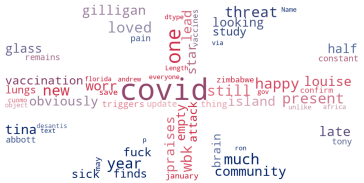



Figure A.24: English word cloud in January


Figure A.25: English word cloud in February


Figure A.26: English word cloud in March

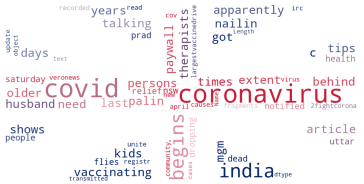



Figure A.27: English word cloud in April


Figure A.28: English word cloud in May


Figure A.29: Spanish word cloud in December


Figure A.30: Spanish word cloud in January

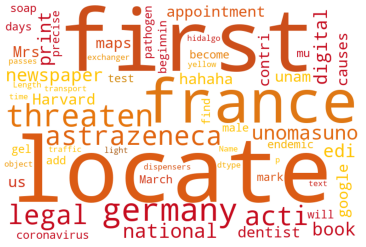



Figure A.31: Spanish word cloud in February

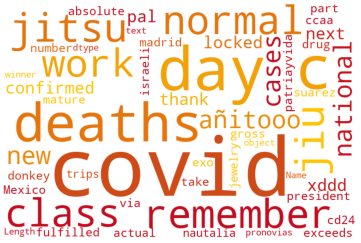



Figure A.32: Spanish word cloud in March


Figure A.33: Spanish word cloud in April


Figure A.34: Spanish word cloud in May


Figure A.35: French word cloud in December


Figure A.36: French word cloud in January


Figure A.37: French word cloud in February


Figure A.38: French word cloud in March


Figure A.39: French word cloud in April


Figure A.40: French word cloud in May

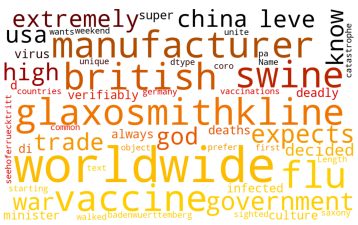



Figure A.41: German word cloud in December


Figure A.42: German word cloud in January

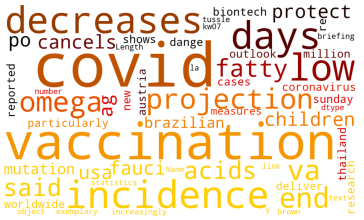



Figure A.43: German word cloud in February

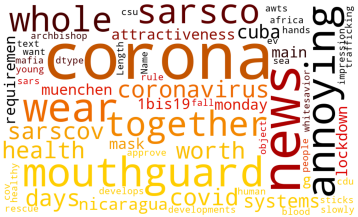



Figure A.44: German word cloud in March

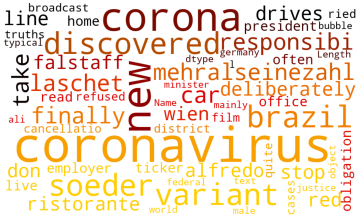



Figure A.45: German word cloud in April


Figure A.46: German word cloud in May

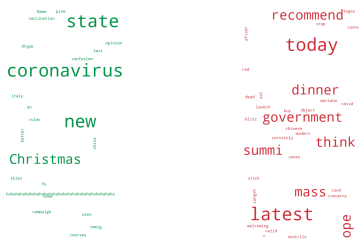



Figure A.47: Italian word cloud in December


Figure A.48: Italian word cloud in January

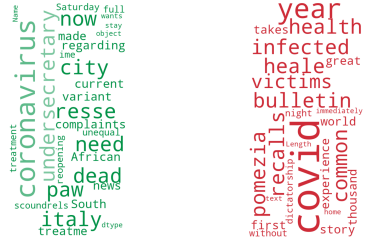



Figure A.49: Italian word cloud in February


Figure A.50: Italian word cloud in March

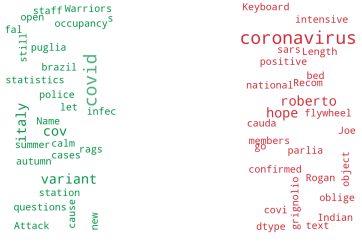



Figure A.51: Italian word cloud in April

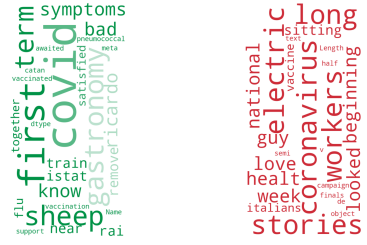



Figure A.52: Italian word cloud in May


Figure A.53: Dutch word cloud in December

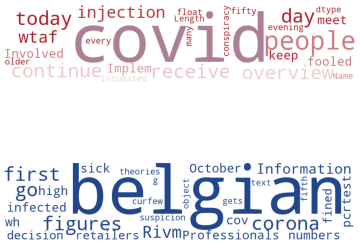



Figure A.54: Dutch word cloud in January


Figure A.55: Dutch word cloud in February


Figure A.56: Dutch word cloud in March

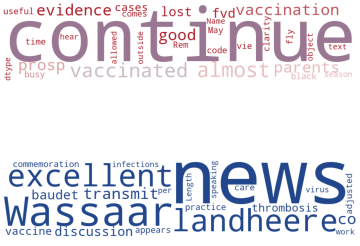



Figure A.57: Dutch word cloud in April

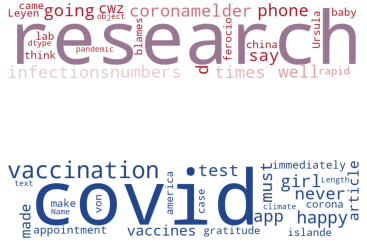



Figure A.58: Dutch word cloud in May

# References


*Public Opinion on Covid-19 Vaccination in The EU 2020*. European Commission, Dec 2020. URL https://ec.europa.eu/info/sites/default/files/covid-19_vaccination_in_the_eu_desk_research_eurobarometer.pdf.

Facebook users in malta - september 2020, 2021. URL https://napoleoncat.com/stats/facebook-users-in-malta/2020/09.

News api – search news and blog articles on the web, 2021a. URL https://newsapi.org/.

Newscatcher news api, 2021b. URL https://newscatcherapi.com/.

Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102, feb 2020. doi: 10.1016/j.ins.2019.10.031.

Alexandra Balahur and Marco Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75, jan 2014. doi: 10.1016/j.csl.2013.03.004.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Katya Artemova, Elena Tutubalin, and Gerardo Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research – an international collaboration, 2021. URL https://arxiv.org/abs/2004.03688.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Jashubhai R. Chaudhary and Joy Paulose. Opinion mining on newspaper headlines using svm and nlp. *International Journal of Electrical and Computer Engineering*, 9(3):2152–2163, 06 2019. Copyright - Copyright IAES Institute of Advanced Engineering and Science Jun 2019; Last updated - 2019-04-01.

W. Christian Crannell, Eric Clark, Chris Jones, Ted A. James, and Jesse Moore. A pattern-matched twitter analysis of US cancer-patient sentiments. *Journal of Surgical Research*, 206(2):536–542, dec 2016. doi: 10.1016/j.jss.2016.06.050.

Nathan Danneman and Richard Heimann. *Social media mining with R*. Packt Publishing Ltd, 2014.

Milagros Fernández-Gavilanes, Tamara Álvarez-López, Jonathan Juncal-Martínez, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58:57–75, oct 2016. doi: 10.1016/j.eswa.2016.03.031.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), May 2014.

Victoria Kayser and Antje Bierwisch. Using twitter for foresight: An opportunity? *Futures*, 84:50–63, nov 2016. doi: 10.1016/j.futures.2016.09.006.

Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4):499–527, aug 2016. doi: 10.1007/s10462-016-9508-4.

N Mamo. Multiplex: A python library to create and annotate beautiful visualizations. *URL: https://github.com/NicholasMamo/multiplex-plot*, 2021.

Joshua Roesslein. Tweepy: Twitter for python! *URL: https://github.com/tweepy/tweepy*, 2020.

H. Tankovska. Twitter: monthly active users worldwide, Jan 2021. URL `https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/`.