Healthcare Analytics 3 (2023) 100172

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A natural language processing approach for analyzing COVID-19 vaccination response in multi-language and geo-localized tweets

Marco Canaparo [a,*], Elisabetta Ronchieri [a,b,*], Leonardo Scarso [c]

[a] INFN-CNAF, Viale Berti Pichat 6/2, Bologna, 40126, Italy
[b] Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, Bologna, Italy
[c] Department of Medical and Surgical Sciences, University of Bologna, Via Pelagio Palagi 9, Bologna, Italy

ARTICLE INFO

Keywords:
Natural language processing
Time series
COVID-19 vaccines
Sentiment analysis
Vaccine hesitancy
Topic modeling

ABSTRACT

Social media platforms, such as Twitter, have been paramount in the COVID-19 context due to their ability to collect public concerns about the COVID-19 vaccination campaign, which has been underway to end the COVID-19 pandemic. This worldwide campaign has heavily relied on the actual willingness of individuals to get vaccinated independently of the language they speak or the country they reside. This study analyzes Twitter posts about Pfizer/BioNTech, Moderna, AstraZeneca/Vaxzevria, and Johnson & Johnson vaccines by considering the most spoken western languages. Tweets were sampled between April 15 and September 15, 2022, after the injections of at least three doses, collecting 9,513,063 posts that contained vaccine-related keywords. To determine the success of vaccination, temporal and sentiment analysis have been conducted, reporting opinion changes over time and their corresponding events whenever possible concerning each vaccine. Furthermore, we have extracted the main topics over languages providing potential bias due to the language-specific dictionary, such as Moderna in Spanish, and grouped them per country. Once performed the pre-processed procedure we worked with 8,343,490 tweets. Our findings show that Pfizer has been the most debated vaccine worldwide, and the main concerns have been the side effects on pregnant women and children and heart diseases.

1. Introduction

Social media platforms, such as Facebook [1], Twitter [2] and Instagram [3], enable people to share their thoughts and their emotional statements about a particular event, from a personal occurrence to a comment on a popular fact [4]. During the COVID-19 pandemic, they have offered relevant information about how people have experienced and perceived the availability of vaccines and the vaccination campaign to reduce the contagion worldwide as well as deaths and hospitalizations [5]. Governments may have prevented the spread of a virus by taking advantage of online posts because they can be extracted and analyzed [6,7].

Vaccination started in December 2020. The entire population has been progressively involved in the injection of different brands of vaccines: Pfizer/BioNTech, Moderna, AstraZeneca/Vaxzevria and Johnson & Johnson. People have been often uncomfortable with the idea of being vaccinated and divided over the usefulness of experimental vaccines, causing a debate that also landed on social media. They constitute precious data sources that offer researchers the chance of extracting data and analyzing knowledge about the COVID-19 vaccination campaign [8–10].

To the best of our knowledge, the majority of previous works on opinions toward COVID-19 vaccination on social media have focused on a single location and a single language [9,11,12]. On the other hand, our study collects posts regardless of location and language.

In this paper, we have extracted and analyzed Twitter posts about Pfizer/BioNTech, Moderna, AstraZeneca/Vaxzevria and Johnson & Johnson vaccines, focusing our attention on the period between April 15 and September 15, 2022, when multiple doses had already been injected and the situation was different from the beginning of the pandemic. The Natural Language Processing (NLP) techniques have been explored to determine the general view about COVID-19 vaccination campaign. We have also tried to link our quantitative outcomes with available news. For a better understanding of our aims we have formulated the following research questions:

RQ1 What have we learned from COVID-19 vaccination in general? It focuses on the acceptance or resistance to vaccination due to the partial availability of full information about the success or failure of vaccination. We have responded through the usage of geo-tagged tweets, topic modeling, and sentiment analysis.

* Corresponding authors.
E-mail addresses: marco.canaparo@cnaf.infn.it (M. Canaparo), elisabetta.ronchieri@cnaf.infn.it (E. Ronchieri).
URL: https://www.unibo.it/sitoweb/elisabetta.ronchieri/ (E. Ronchieri).

https://doi.org/10.1016/j.health.2023.100172
Received 25 February 2023; Received in revised form 27 March 2023; Accepted 5 April 2023

RQ2 What is the impact of language on tweet-based research about COVID-19 vaccines? It highlights how languages may introduce a bias in this kind of tweet-based research. We have used topic modeling and time series.

RQ3 Which brands of COVID-19 vaccine have been the most debated in different countries and languages? We have used hierarchical clustering and statistical techniques.

RQ4 What have been the main language-specific topics of discussion regarding COVID-19 vaccines? We have used topic modeling.

Our findings show that people's reactions to events all over the world seem to follow similar patterns. More in detail, bad news, which sometimes turns out to be fake, seems to spread faster and generate a larger discussion than good news, as confirmed also in previous literature [13]. These observations have been possible by using geo-tagged tweets [2,11,14] and considering the main languages found in the collected tweets. The Pfizer brand has been shown to be the most popular debated vaccine all over the world. Furthermore, concerns for children, pregnant women, and heart diseases have been highlighted.

2. Related works

In the past three years, a large number of studies has surfaced in the literature, investigating the potential role of social media in exploring individuals' opinions towards COVID-19 vaccination, as well as in tackling vaccine hesitancy. In the following, a list of studies that is pertinent to our research is provided.

2.1. Independent on languages and locations

Umair et al. [15] highlight the importance of vaccine development to control the COVID-19 pandemic and the crucial role of vaccine distribution in developing immunity against the virus, and analyze public opinion about COVID-19 vaccines using social media platforms.

Qorib et al. [16] investigate COVID-19 vaccine hesitancy using three different sentiment computation methods (Azure Machine Learning, VADER, and TextBlob). and ML algorithms with different combinations of three vectorization methods (Doc2Vec, CountVectorizer, and TF-IDF).

Yousefinaghani et al. [17] collect 4,552,652 publicly available tweets between January 2020 and January 2021, identifying the vaccine sentiments and opinions and comparing their temporal progression, geographic distribution, main themes, keywords, post engagement metrics, and account characteristics.

Andreadis et al. [14] have proposed a novel framework that collects, analyses, and visualizes Twitter posts in order to predict their reliability and detect trending topics and events.

Puri et al. [18] examine digital health tactics for managing vaccine misinformation on social media, including message framing and utilizing public figures.

Chou and Bundez et al. [19] document that individuals have strong emotional responses to COVID-19 vaccines, and stress the significance of emotion in vaccine communication strategies aimed at tackling vaccine hesitancy.

2.2. Language and/or location specific

Sussman et al. [20] collect social media COVID-19 vaccine-English mentions, and examine topics and sentiments during the period of September 1, 2020, through December 31, 2020.

Lanyi et al. [2] design a platform to analyze barriers to vaccines, focusing on geo-located tweets from London, UK, from 30th November 2020 to 15th August 2021, identifying safety concerns, accessibility issues, and misinformation as the most common factors that hinder vaccine uptake.

Jang et al. [9] aim to investigate users' attitudes toward COVID-19 vaccination in Canada after vaccine rollout.

Ogbuokiri et al. [11] extract and cluster Twitter posts to study variations in sentiments toward COVID-19 vaccine related topics, focusing on the geolocated tweets coming from the three largest South African cities (Cape Town, Durban, and Johannesburg) and detecting vaccine uptake as the most discussed topic.

Aleksandric et al. [21] investigate the correlation between social media sentiments related to COVID-19 vaccination and their impact on vaccination rates in the United States (US), by considering English tweets gathered between January 4th and May 11th, 2021.

Mishra et al. [22] conduct a comparative analysis of public sentiments towards COVID vaccination in India before and after the second wave, comprising of 5977 tweets before the second wave and 42,936 tweets after the second wave.

Ljajic et al. [23] identify the reasons for vaccine hesitancy in negative tweets written in Serbian, employing NLP techniques and classifying tweets related to vaccination based on their sentiment polarity.

Kwok et al. [24] collect and analyze 31,100 English tweets containing COVID-19 vaccine-related keywords from Australian Twitter users between January and October 2020.

Ahmed et al. [25] establish a causal relationship between the use of social media for news consumption and vaccine hesitancy in the US population, finding that the utilization of social media as a source of news content may contribute to vaccine hesitancy and amplify citizens' doubts about the effectiveness of vaccines.

Ogbuokiri et al. [26] gathered 70,000 geotagged vaccine-related tweets across nine African countries from December 2020 to February 2022, categorizing them into positive, negative, and neutral sentiments. The findings of this study demonstrate that social media data can be utilized to supplement current data in detecting outbreak hotspots and to address vaccine hesitancy.

Martínez et al. [27] introduce a newly annotated corpus of 2801 tweets related to COVID-19 vaccination, which was annotated by three native Spanish speakers as being in favor (904), against (674), or neutral (1223) with a Fleiss' kappa score of 0.725.

Lindelöf et al. [28] examine 16,713,238 English tweets related to COVID-19 vaccines, covering the period from March 1, 2020, to July 31, 2021, with a particular focus on those expressing negative attitudes toward vaccination in order to investigate the evolution of negative tweets and to explore the various topics discussed in the tweets.

Biancovilli et al. [13] analyze the Brazilian case of COVID-19 *infodemic risk*, reporting the main sources and the type of misinformation, as also the topics that circulate the most, demonstrating that the posted misleading information reflected in the high number of COVID-19 cases that the country had from the beginning of the pandemic to May 2021.

Alsudias et al. [29] analyze Arabic tweets that can contain informal terms to assess their usefulness for health surveillance, identifying the locations where the infection is spreading and employing deep learning techniques in the classification process.

Gori et al. [30] analyze the main COVID-19 vaccination events in Italy by using the relative increase indicator ΔT, which measures the ratio between the difference between the number of tweets on a specific day and the number of tweets of the day before (ΔT_{day} - ΔT_{day-1}), divided by the weekly amount of tweets (ΔT_{week}): a high value of the relative increase indicator on a day indicates a high probability that the related event has raised an anti-Covid-19 discussion.

Griffith et al. [31] conduct a content analysis of 3915 tweets from public Twitter profiles in Canada, using the search terms "vaccine" and "COVID", to identify tweets related to COVID-19 vaccine hesitancy, classified according to the inclusion criteria and organized into themes using content analysis.

Fazel et al. [32] analyze approximately 2 million tweets from 522,893 individuals in the United Kingdom between November 2020 and January 2021 to investigate the associations between Twitter discourse on vaccines and significant scientific news releases related to vaccines.

2.3. Focus on vaccine brands

Marcec et al. [33] make innovative research on how daily average sentiment analysis could be different for each COVID-19 injected vaccine, gathering information when there are drastic changes in the sentiment scores. They have proved that Pfizer and Moderna vaccine sentiment scores have been constantly positive during December 2020 and 2021, while the AstraZeneca vaccine has significantly decreased over time.

Huangfu et al. [8] analyze the overall sentiments and topics by exploiting 1,122,139 tweets about COVID-19 vaccines (Pfizer, Moderna, AstraZeneca, and Johnson & Johnson) collected from December 14, 2020, to April 30, 2021.

2.4. Focus on dataset or framework construction

Chen et al. [34] focus on the construction of a dataset of 2,198,090 tweets related to COVID-19 vaccines, collected from Western Europe and annotated by the authors to extract vaccination attitudes from these social media posts.

Sepúlveda et al. [35] introduce COVIDSensing, a Web-based tool that creates new time series as a combination of different sources of information (i.e., Telegram) and can predict the evolution of the pandemic by considering epidemiological and clinical variables.

Banda et al. [36] explore the necessary cleaning operations to be performed on a raw dataset that has become a publicly available versioned dataset: the less recent version contains the raw texts obtained via the Twitter Stream API; the more recent one includes the results obtained through the Social Media Mining Toolkit.

Guntuku et al. [37] employ Twitter to obtain a randomized subset of more than 78 million vaccine-related tweets, published between December 1, 2020 and February 28, 2021 with the aim of elucidating the temporal and geographical variation in the discourse surrounding COVID-19 vaccines.

Cornelius et al. [6] construct the COVID-19 Twitter Monitor, which is the result of several NLP techniques applied to social media posts, offering a new way to access information through this interactive platform.

3. Research methodology

Machine learning and natural language processing techniques are used to analyze tweets in order to identify barriers through the COVID-19 vaccines after the third vaccine dose roll-out.

3.1. Data collection strategy

Twitter is a social media platform that has been used as a source of real-time user-generated data to track public perceptions over time [2]. It allows academic researchers to define projects for which they can make API requests to extract data for their analysis purpose. Fig. 1 explains the entire process of data collection that we have performed through a Twitter Stream API [38], called *Tweepy* [39], which downloads Twitter messages under the authorization of Twitter Developer. This API connects to social media platforms and extracts data according to a parameterized search strategy. It is based on the proper vaccine names and some other terms, e.g. the cardinal number of injected doses (such as first dose, and second dose).

Table 1 shows the search terms used to gather COVID-19 vaccination-related tweets. Our query contains the bi-grams COVID vaccine and COVID vaccination and their respective plurals, along with the combinations involving COVID-19. Furthermore, the query incorporates the proprietary names of vaccines (i.e., Pfizer, Moderna, AstraZeneca and Johnson & Johnson), specific names (i.e., BioNTech, and VaxZevria), and the dose-related words, mainly preceded by the injection number (i.e., the first dose, second dose, third dose, and

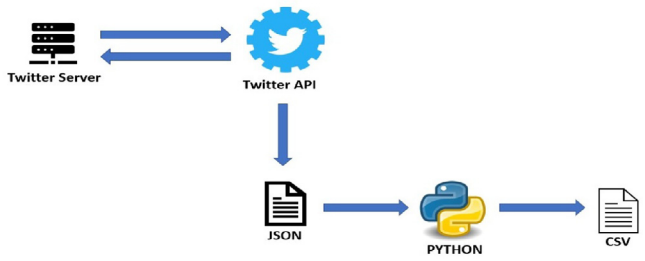

Fig. 1. Data extraction.

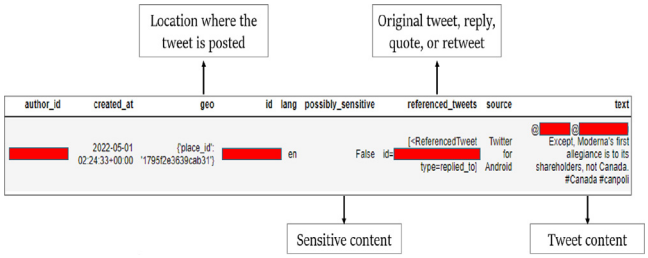

Fig. 2. Example of attributes derived through Tweepy.

Table 1
Search strategy terms.

Keywords
Pfizer, Pfizer BioNTech, BioNTech, Moderna, AstraZeneca, Johnson & Johnson, J & J, VaxZevria
first dose, second dose, third dose, fourth dose
COVID{-19} vaccine{s}, COVID{-19} vaccination{s}
prima dose, seconda dose, terza dose, quarta dose
COVID{-19} vaccin{o,i}, COVID{-19} vaccinazion{e,i}

fourth dose) in both English and Italian languages. The rationale behind incorporating dose-related keywords in the Italian language despite the bulk of collected tweets being in languages other than Italian was to focus on the first European country that faced COVID-19. We have collected tweets on a daily basis, from April 15 to September 15, 2022. The specific search terms have been generated through discussion among the authors of this paper, along with checking relevant literature [2,33,40].

The data have been collected over a five-month period between April 15 and September 15, 2022. This time period covers the spring and summer seasons after the third vaccine dose roll-out. Fig. 2 shows the set of selected attributes with their values that have been saved in a comma-separated variable (csv) file.

3.2. Data pre-processing and geo-location

The pre-processing phase aims at producing a clean tweet message [41]. To achieve this purpose, raw tweets have undergone some elaboration activities: firstly, the text has been made lowercase; secondly, we have filtered out some characters, such as the hyphen, and the ampersand & symbols; furthermore, we have removed the usernames to protect the privacy of users, we have canceled the punctuation, the special characters as well as the non-ASCII characters; finally, we have filtered out the stop words according to each language of the dataset.

Table 2 shows tweet examples whose text has been pre-processed. In the first example, the retweet symbol 'RT' has been eliminated, as the punctuation marks and the username, that it is dropped for privacy policy; moreover, the text has been made lowercase, since the following analysis could be case-sensitive. The second example shows the way the

Table 2
Example of text cleaning of COVID-19 related tweets.

Before	After
RT @username COVID-19 Vaccination is important!!	covid 19 vaccination important
Today I had my THIRD dose \n #Pfizer :-)	today third dose pfizer


```
{
  "id": "df51dec6f4ee2b2c",
  "url": "https://api.twitter.com/1.1/geo/id/df51dec6f4ee2b2c.json",
  "place_type": "neighborhood",
  "name": "Presidio",
  "full_name": "Presidio, San Francisco",
  "country_code": "US",
  "country": "United States",
  "contained_within": [
    {...
```


Fig. 3. Example of a JSON object of a geo-tagged tweet.

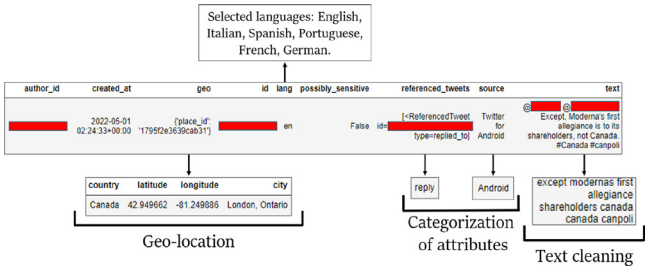

Fig. 4. Example of attribute transformation after step 2.

emoticons have been treated, so long as they have been removed, as the hashtag symbol '#'.

Afterward, we have kept the ones written in the six more recurring languages in our collection, such as English, French, Spanish, Portuguese, German, and Italian.

The geo-location phase consists of processing the geographic information in the geo-tagged tweets. They are a subset of all the collected tweets characterized by the presence of the users' location sharing. This kind of tweets enables us to geographically localize reactions to an event all over the world. All geo-location data have been transformed into JavaScript Object Notation (JSON) format. Fig. 3 shows an example of a geo-tagged tweet: the last three attributes contain the user's geographic information and their acronyms stand for country name, city name, and latitude and longitude respectively. We have derived the continent feature by using the country_code attribute, whose value follows the ISO 3166-1 alpha-2 format [42], and replaced it with the continent name.

Fig. 4 shows the same tweet of Fig. 2 transformed according to the second step.

3.3. Qualitative analysis

After data pre-processing and geo-location, a qualitative analysis is presented to investigate into the dominant debate about COVID-19 vaccination.

Vaccine occurrences in tweets give us insights into the daily importance of vaccination over the general discussion. For this purpose, we have defined two measurements:

1. Absolute count of vaccines (ACV) that counts the different types of vaccines, i.e. Pfizer, Moderna, AstraZeneca, Johnson & Johnson (i.g. the ampersand & has been removed in the pre-processing phase).
2. Relative count of vaccines (RCV) that is the ratio between ACV and the total number of daily tweets.

We can mathematically represent RCV through the following expression:

$$RCV_{i,j} = \frac{ACV_{i,j}}{T_j} \tag{1}$$

where $ACV_{i,j}$ represents the absolute count of the vaccine ith on tweets of the day jth and T_j is the number of collected tweets (T) on the day jth. For example, we can interpret a value of the $\text{RCV}_{Pfizer,j}$ of 0.89 as the Pfizer vaccine is presented once in the 89% of the tweets collected on day jth.

ACVs and RCVs trends have been visualized during the COVID-19 vaccination campaign with respect to different languages [43] and related to events that happened in the same period of time. Furthermore, geo-location information has been exploited to group countries with similar RCVs and consequently with comparable discussions present in tweets about the COVID-19 vaccination. Finally, RCV values over different languages and continents have been compared to give another intuitive representation of the COVID-19 vaccine discussion.

3.4. Keyword trend analysis

At this stage of our study, the pre-processed tweets have been parsed into individual terms, which form our corpus. From this, a dictionary comprising all the unique words can be generated. We have used the count vectorizer and term frequency-inverse document frequency (TF-IDF) for feature extraction to determine first the most mentioned words [44] for each language and later the main topics.

The count vectorizer technique converts given tweets into a vector space and covers words that frequently occur in the tweets. This technique creates a word matrix where every unique word represents the column of the matrix and the selected tweet from the dataset represents the row of the matrix. We count the word in that particular set of tweets.

The TF-IDF technique is used along with the count vectorizer. TF-IDF (Eq. (2)) takes the TF (Eq. (3)) part and its corresponding IDF (Eq. (4)) as a product tth term and dth document to get the weights of features in a document:

$$TF - IDF = TF_{t,d} * IDF_t \tag{2}$$

$$TF_{t,d} = \frac{N_{t,d}}{N_d} \tag{3}$$

$$IDF_t = \log \frac{\text{N}}{1+df_t} \tag{4}$$

where the number of $TF_{t,d}$ is denoted by the ratio between the number of times the tth term appears in the dth document ($N_{t,d}$) and the total number of terms into the dth document (N_d). The number IDF_t is the inverse of the document frequency for each term, computed by dividing the total number of documents in a corpus (N) and the document frequency for each tth term (df) plus one, on a logarithmic scale. When a term t frequently appears in many documents, IDF computes the weights of a phrase t low. For example, stop words have a lower IDF value.

To analyze the main topic across our dataset, we have applied the Latent Dirichlet Allocation (LDA) topic model [45] to the TF-IDF features. The LDA model assures that the estimated number of topics is created, and each theme contains a list of words and its corresponding weight. The number of topics is estimated by adopting the coherence score [46], which is a measure of the adaptability of the model to

Table 3
Frequency distribution of tweets with respect to the absence/presence of sensitive content.

possibly_sensitive	Number of tweets	Percentage of tweets
false	9,408,993	98.9%
true	104,070	1.1%

the created dictionary. Each theme is afterward interpreted by the researchers [47].

Moreover, to identify, extract, quantify, and study the opinion about a text, we have computed the polarity of a given document, which quantifies an emotional statement (e.g., anger, pain, fear, and happiness), and a positive, negative, or neutral classification that can be assigned to the sentence. In this case study, a free open-source library, *spaCy* [48] has been used that supports a wide range of languages. The model assigns a polarity between −1.0 and 1.0, which respectively indicates the most negative score and the most positive score; the original content of the observations has been considered since it has been proved that the emoticons (e.g. :) that represents the smile) change significantly the score of a tweet. The pre-processed text does not contain any punctuation marks. The results that will be reported in the following are the averaged daily scores for each language, with the aim of detecting the reasons why emotional statements about vaccination campaigns change over time.

4. Pre-processing and geo-location results

4.1. Frequency distributions

The obtained dataset contains a total of 9,513,063 observations. Each observation is characterized by the following attributes: author_id is the identifier of the tweet author; created_at is the timestamp when the tweet is posted; geo contains the information about the location where the tweet is posted; id is the numerical identifier of the status; lang is the language in which the tweet has been written; possibly_sensitive is a variable indicating if the tweet contains sensitive content according to the guidelines; referenced_tweets indicates if a tweet is original, a reply, or a quote, followed by the tweet identifier from which it originated (in the case the tweet was not original); source is the name of the app the user tweeted from; text includes the message content posted by the user.

With the collected data, we have filtered out all the tweets that have a true value in the possibly_sensitive variable (see Table 3), which indicates that a tweet content may be recognized as sensitive according to Twitter guidelines. Furthermore, we have removed the minor language tweets (12.29% of the observations) leaving a total of 8,343,490 tweets, as shown in Table 4. This last operation is essential since the next phases involve language information.

The source attribute has 4 values: Web App, Apple, Android, and Other according to the platform with which the tweet has been posted. Table 5 shows that the majority of the tweets contain the Android value, followed by Apple, Web App, and Other.

We have also identified our referenced_tweets values, such as original tweet, which indicates a message posted for the first time; retweet, which indicates a message shared by another user (maybe followed by a comment); quote, which means a tweet similar to a retweet, displaying a reaction message tacked under the original message; reply, which indicates a comment posted under a previously written tweet. The variable containing the previous four values has been renamed as type_of_tweet. The most frequent category is retweet, and the least frequent category is quote, as shown in Table 6.

Table 4
Frequency distribution of tweets with respect to their language.

Lang	Number of tweets	Percentage of tweets
English	5,341,026	64.01%
French	1,175,212	14.09%
Spanish	960,008	11.51%
Italian	346,343	4.15%
Portuguese	276,216	3.31%
German	244,685	2.93%

Table 5
Frequency distribution of tweets with respect to their source.

Source	Number of tweets	Percentage of tweets
Android	3,200,144	38.35%
Apple	2,671,202	32.02%
Web App	2,204,423	26.42%
Other	267,721	3.21%

Table 6
Frequency distribution of tweets with respect to their type.

type_ of_tweet	Number of tweets	Percentage of tweets
Retweet	6,533,088	78.03%
Reply	957,087	11.47%
Original	738,615	8.85%
Quote	114,700	1.37%

Table 7
Frequency distribution of geo-tagged tweets with respect to their continent.

Continent	Number of tweets	Percentage of tweets
North America	9,149	40.46%
South America	5,736	25.36%
Europe	4,916	21.74%
Asia	1,571	6.95%
Oceania	750	3.32%
Africa	492	2.17%

4.2. Distribution of geo-tagged tweets

Geo-tagged tweets, as described in Section 3.2, are used to analyze the distribution of COVID-19 tweets all over the world, because they contain the exact location of the city where the user posted the message. They are 22,614, thus 0.27% of the total number of collected tweets. From the city field, we have identified geo-location data, such as coordinates, country, and continent. Table 7 shows the frequency distribution of these tweets.

Geo-location data have been used to create an interactive globe map (statically shown in Fig. 5 built by using the *Folium* python library [49]) showing how tweets related to COVID-19 vaccinations are spread across the world. This map shows as many circle markers as the geo-tagged tweets. The number inside a circle has been represented in three different colors: red, yellow, and green, respectively depending on the magnitude of geo-tagged tweets within the region below. A darker color of the marker entails a larger number of geo-located tweets in that area. The green pop-ups on each circle marker contain the country and city names of the tweet.

5. Time series results

In this section, we analyze the time series plots on ACV and RCV values (defined in Section 3.3) for the various vaccines and languages to identify similarities and differences for each selected language. The plots are representations of the indices over time. Our aim is to speculate about a possible correlation between a peak of ACV and RCV for a certain type of vaccine and the report of important events all over the world.

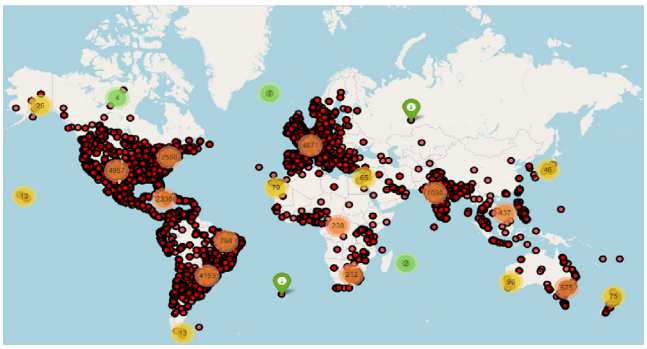

Fig. 5. Geographic map of geo-tagged tweets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

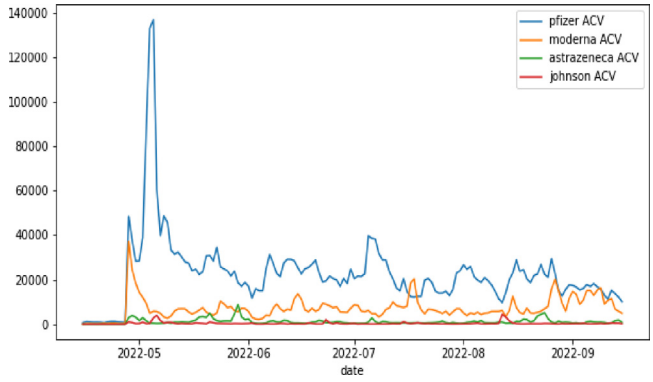

Fig. 6. Time series of the ACVs of COVID-19 related tweets.

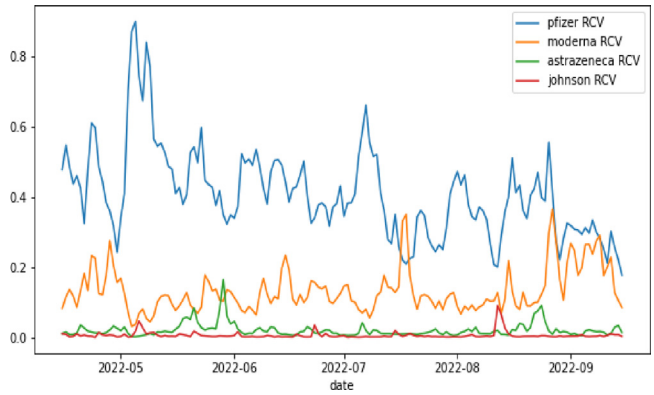

Fig. 7. Time series of the RCVs of COVID-19 related tweets.

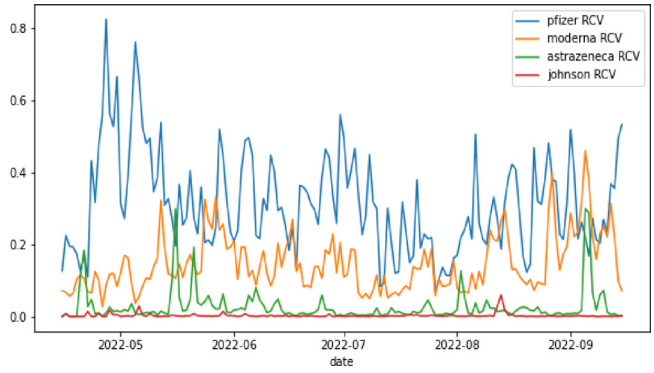

Fig. 8. Time series of COVID-19 related tweets written in the Italian language. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 6 shows the absolute counts of each vaccine. From April 15 to April 27, 2022, the four lines are near the zero value; this is because the collected tweets have been about 3000 per day. In the following days, more than 60,000 tweets are the daily average number of collected tweets, and all four ACVs increase a lot in magnitude. From the second half of May, the plots remain stable until the end of August, after that there is a slight decrease. The highest peak for the Pfizer vaccine appears in the first decade of May when more than 100,000 tweets have been daily collected, and the word Pfizer (or BioNTech) is mentioned more than 140,000 times in a day across the world. Moderna has its first peak during the first days of May, which shows an ACV of about 40,000. The ACV index conveys an absolute value that is not related to the total tweets present across time. As a consequence, we have introduced the RCV index.

The construction of the RCV index allows the comparison of the four series. Fig. 7 shows that Pfizer is the most mentioned vaccine of all time, except at the end of April and half of July, when the Moderna time series has a higher magnitude. The lowest value has been reached on September 15, 2022, when $\mathrm{RCV}_{Pfizer,15/09}$ is equal to 0.178; this means that the word Pfizer compares once in 17.8% of the tweets collected on September 15, 2022. On the contrary, the highest value of this vaccine is 0.90 on May 5, 2022.

In the following, we have performed a deeper analysis based on the different languages of our tweets.

Fig. 8 shows the time series of RCV data for the Italian language. The Pfizer RCV in the blue line is the one with the highest values despite some exceptions: on April 21, 2022, AstraZeneca has reached an RCV value of 0.18 against the 0.17 of Pfizer; on May 16, 2022, the AstraZeneca RCV has been 0.298; from May 25 to May 27, 2022, Moderna has reached a value varying between 0.24 and 0.33, overtaking all the other counts; finally, on September 5, 2022, Moderna has reported a value of 0.45. Pfizer RCVs from April 26 to April 28, 2022, are about 0.8, two possible explanations of these high values might be the discussion following two announcements: one about the non-necessity of having the fourth dose of the Pfizer vaccine and the other about the Pfizer permanent authorization. A value of 0.76 has been reached on May 5, 2022, when the discussion was focused on the reports about the Pfizer's efficacy in the 7 days immediately after its injection. On September 1, 2022, there is a peak of 0.51 related to hundreds of retweets of the news about the authorization from the FDA (the Food and Drug Administration, USA) of the updated Pfizer and Moderna boosters. Regarding Moderna, the $\mathrm{RCV}_{Moderna,25-27/05}$ highlights the spread of a tweet with a photo about an adverse reaction to the second dose of Moderna, showing an arm with a lot of red points on it. $\mathrm{RCV}_{AstraZeneca,21/05}$ of 0.19 shows the discussion about a correlation between the presence of an adenovirus of chimpanzees contained in the vaccine and the development of monkeypox. The consequences of the mandatory vaccine in young people has been discussed on August 2, 2022, when the RCV is 0.12.

Fig. 9 shows the time series of RCV data for the English language and its plots follow a trend similar to Fig. 7, which shows the RCV values of all tweets. This similarity is due to the influence of English tweets on the total number of tweets; they constitute 64.01% of the total. The Pfizer series (the blue one) is above the other ones almost every time except for a few Moderna peaks: e.g. the 0.25 one is between April 28 and April 29, 2022, when there were a lot of comments about a possible injection of this vaccine on people aged from 6 months to 6 years; the 0.21 one on August 15, 2022, is related to the announcement of the approval of the United Kingdom's first bivalent COVID-19 booster vaccine, made by the pharmaceutical company. Fig. 9 also shows a similarity with Fig. 8 with respect to the peak of the Pfizer vaccine on May 5, 2022. A value of 0.65 on July 7, 2022, regards the withdrawal

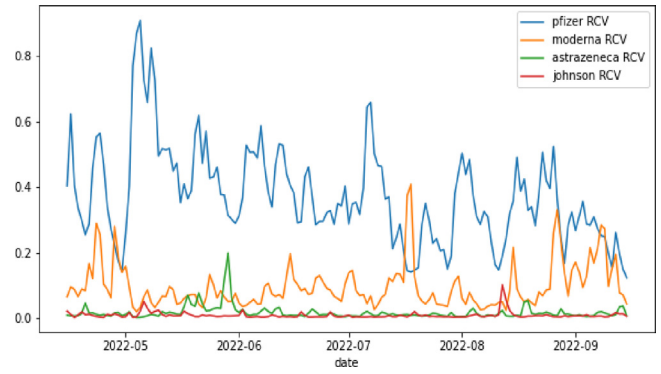

Fig. 9. Time series of COVID-19 related tweets written in the English language. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

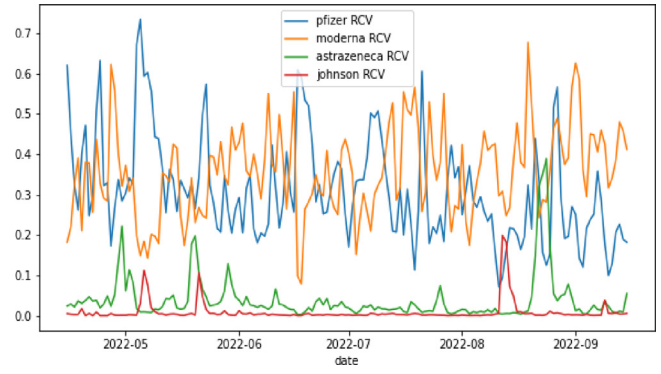

Fig. 10. Time series of COVID-19 related tweets written in the Spanish language. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the Pfizer vaccine from Uruguay. The AstraZeneca series (the green one) is constantly near the 0 value: this may be the consequence of the fact that a lot of English-spoken countries did not inject the vaccine. It is important to remind us that the AstraZeneca vaccine was mostly used in European countries. However, at the end of May, the series has a value of about 0.2, because of a discussion on the possible neurological consequences of vaccine injection.

Fig. 10 shows the Spanish time series that has wide swings in their trends. This might be due to the fact that Spanish-writing people are spread across the world, and so the collection of related tweets takes into account a large number of events from several countries. As a matter of fact, Pfizer and Moderna series (respectively, the blue and yellow ones), interchange a lot. From the second half of August, also AstraZeneca and Johnson & Johnson join the trend of the series. Starting from the Pfizer vaccine, a peak of magnitude 0.75 is reached on May 5, 2022, for the same reason as in the English time series. A peak of 0.5 registered on July 7, 2022, regards the discussion about the correlation between vaccine inoculation and cases of heart failure. The Moderna series does not allow us to detect events related to the vaccine, since the moderna term means modern in the Spanish language. Three main peaks characterize the green line, related to AstraZeneca vaccine. On April 30, 2022, with an RCV of 0.20, a tweet about a person who allegedly died because of the vaccine became viral. The second peak of May 20, 2022, regards a complaint from a Spanish association. The last peak of 0.37, reached between August 22 and August 24, 2022, is probably due to the spread of memes and discussions after the withdrawal of the vaccine.

Fig. 11 shows the time series of tweets written in the Portuguese language. It is characterized by a fast trend change, depending on the

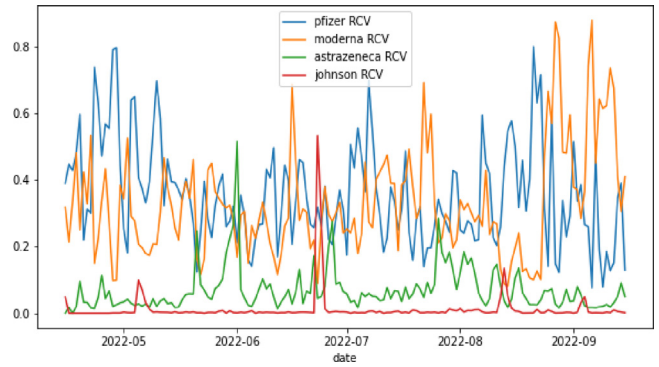

Fig. 11. Time series of COVID-19 related tweets written in the Portuguese language. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

number of collected tweets that are a lower number than the English ones (as shown in Table 4). As in the Spanish case, the Moderna series (the yellow line) is generated by a lot of tweets containing the moderna term that is also used in different contexts. However, on June 16, 2022, the RCV value is higher than 0.60, given by the definitive approval of the vaccine. Regarding the AstraZeneca vaccine, the $\mathrm{RCV}_{AstraZeneca,01/06}$ index is caused by the fast spread of a tweet related to the presence of the "Sale under prescription" sentence in the vaccine package. For the Pfizer time series (the blue line), the safety and the effectiveness of the vaccine are mostly commented on April 29, 2022, since the RCV reached a magnitude of 0.8, and on May 10, 2022, the discussion is about the injection of the vaccine in pregnant women. At the beginning of July, Twitter users focused their attention on some adverse reactions that happened after the injection of the booster dose.

The behavior of the French time series is different from the ones analyzed until now. Fig. 12 shows that the Pfizer line is higher than the other ones over time, leading to think that the French Twitter population focused on the American pharmaceutical company vaccine. Surprisingly, during the first decade of May (the days from May 4 to May 9, 2022) the RCV of the Pfizer vaccine reached a value of 1.07, meaning that the Pfizer term appeared more than once on each tweet. This could be the result of the usage of these words in both text and hashtags. A lower but still significant value of 0.8 is given on May 23, 2022, due to a tweet about alleged vaccine-correlated complications. Two more peaks characterized the month of June: on June 8, 2022, after the news about the injection of a physiological serum instead of the Pfizer vaccine; on June 20, 2022, after the closing of the deal between India and the vaccine company. Other two peaks have been recorded for the Pfizer series in July: one on July 7, 2022, with a magnitude of 0.86, because of the retweets of the news reporting some vaccine-related complications; the other peak of 0.7 is shown on July 29, 2022, due to a video about the vaccine-related consequences on young children and a discussion about vaccine exemption. The highest value obtained from the Moderna series is more than 0.4 on June 14, 2022, when the FDA authorized the usage of the vaccine in 6 months through 17 years' children. For the AstraZeneca vaccine, on July 6, 2022, a peak of 0.18 regards a study of April 2022 about the adverse reactions to the vaccine. For Johnson & Johnson plot (the red one), no particular event affected the discussions of French people.

The last analyzed language is German, plotted in Fig. 13. Similar to the French case, the Pfizer line overtakes the other series, with a value of 1 obtained on May 4, 2022, and May 5, 2022. The causes of these peaks seem to be similar to the ones of the previous languages. The same magnitude is given on May 20, 2022. The Moderna peak on May 18, 2022, may be located in Switzerland since it regards the approval documents of the Moderna vaccine in that country. On June 22, 2022, a peak in the Moderna time series might be due to alleged vaccine-related consequences.

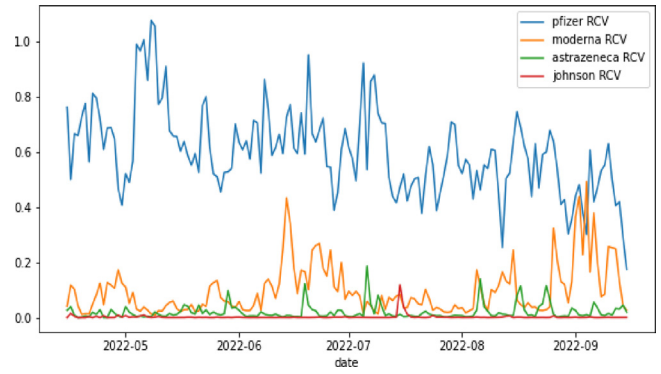

Fig. 12. Time series of COVID-19 related tweets written in the French language. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

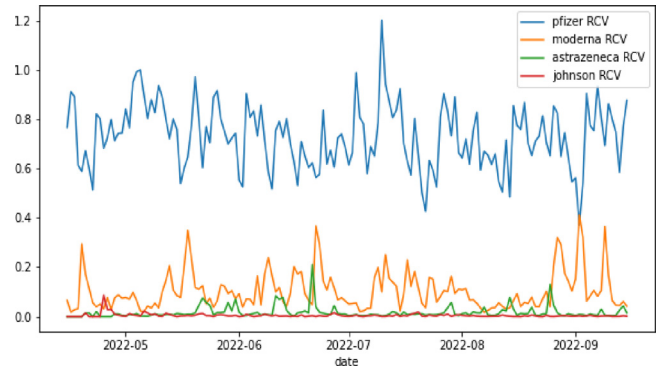

Fig. 13. Time series of COVID-19 related tweets written in German language.

6. Data exploration by hierarchical clustering

We show our results on the comparison among the RCV values achieved by different countries. Our purpose is to group countries that showed a comparable discussion about COVID-19 vaccination. We have used the Hierarchical Clustering [50] method that groups data objects into hierarchies or "trees" of clusters. More specifically, we have employed Agglomerative Hierarchical Clustering [51] that starts by letting each object form its own cluster and iteratively merges the cluster into larger ones. This clustering technique has been used in combination with geo-tagged tweets and has allowed us to build a dendrogram [52] that organizes clusters according to their distance in terms of RCV values. We have made a hypothesis about how many clusters are necessary: we have chosen to organize results into four clusters because we have categorized each vaccine discussion as very strong (VS), strong (S), weak (W), and very weak (VW). The applied criterion to build the hierarchical clustering is based on the Ward's method [53] that relies on an optimal value of an objective function, and, at each step, it merges a pair of clusters according to their distance, to which a cutoff value is assigned. Based on the K-means objective function, the Agglomerative Cluster technique enables the merging of cluster pairs with the smallest possible value, i.e., the two most similar groups in terms of RCV value at each iteration. This method incrementally improves the dissimilarity value while aiming to maintain cluster homogeneity [54]. Unlike traditional partitional clustering techniques such as Partitioning Around Medoids, the Agglomerative Hierarchical Clustering approach provides a more meaningful and subjective way of dividing clusters. This is because the Agglomerative technique uses a dissimilarity measure based on the data structure, that is, in our case, the Ward's method [55]. Consequently, this analysis may lead to group countries with similar RCVs, meaning that people of these countries have similar feelings towards the COVID-19 vaccination.

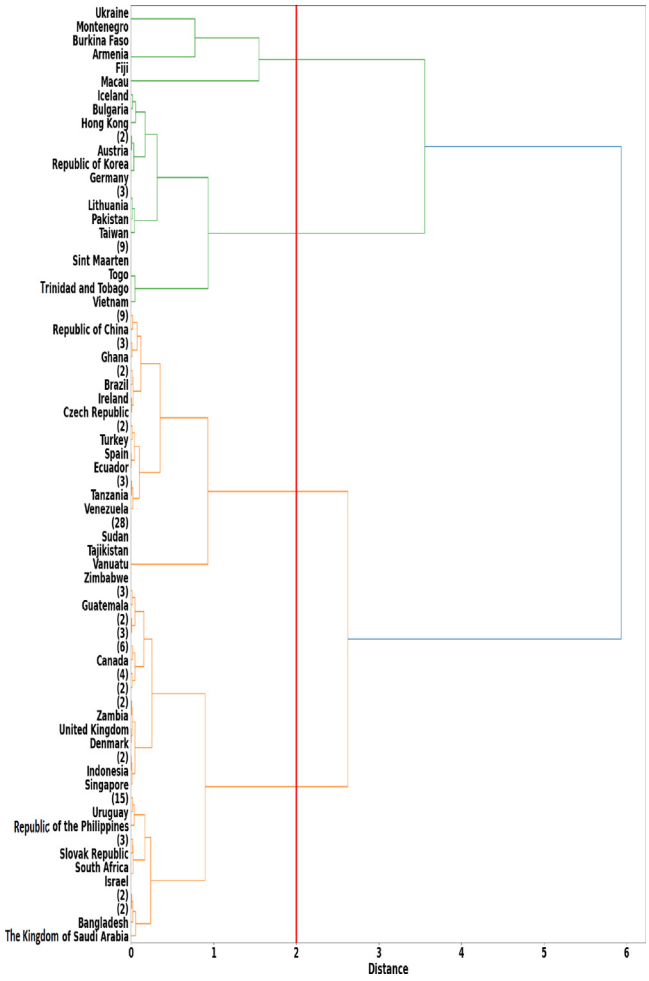

Fig. 14. Hierarchical clustering of the countries of geo-tagged tweets with respect to their Pfizer RCV.

Fig. 14 shows the way the clusters have been obtained for the Pfizer RCV vaccine. The red vertical line (i.e. the cutoff line), at a distance equal to 2, cuts the dendrogram into four clusters. On the left side, there is a list of countries that made part of the clusters: the numbers in round brackets represent the amount of countries grouped at a minimum variance increase level; indeed, for display purposes, the entire list of countries is not shown. Starting from the top of Fig. 14, the first obtained cluster consists of Ukraine, Montenegro, Fiji, Armenia, Macau and Burkina Faso. They are similar since the corresponding RCV_{Pfizer} for these countries has a value between 1.5 and 3. Each geo-tagged tweet of the first cluster contains the words Pfizer/Biontech at least more than once, even three times (as in the case of Macau), leading to categorize these countries as the ones with a very strong discussion. Table 8 shows some examples of countries for each RCV category of the Pfizer vaccine. The second cluster contains 27 countries spread across the world, e.g. Trinidad and Tobago in America, Germany in Europe, Taiwan in Asia, and Guinea in Africa. The cluster corresponding to the weak class of the Pfizer-related discussion contains 60 countries. In addition, many geo-tagged tweets from Asia (such as Russia, Thailand, Israel, and Lebanon) contain a moderate discussion of the vaccine; most of the North African and South African countries take part in this cluster. The very weak category contains 61 countries, some of which are Turkey, Hungary, India, and most of the smallest countries spread across the world: from the ones of western Asia to the ones of central Africa, to the islands of North America and Oceania. The relative RCV ranges from 0.25 to 0, when the geo-tagged tweets do

Table 8
Category discussion over Pfizer RCV values.

Category discussion	RCV range	Countries
VS	[1.5, 3]	Ukraine, Monte Negro, Fiji, Armenia, Macau, Burkina Faso
S	[0.67, 1]	Trinidad and Tobago, Nicaragua, Sint Marteen, Germany, Austria, France, Croatia, Taiwan, Pakistan, Republic of Korea, Hong Kong, Togo
W	[0.27, 0.6]	United Kingdom, Ireland, Czec Republic, Spain, Ghana
VW	[0, 0.25]	Bangladesh

Table 9
Category discussion over Moderna RCV values.

Category discussion	RCV range	Countries
VS	[1, 2]	Georgia, Niger, Sudan, Vanuatu and Burkina Fasu
S	[0.35, 0.8]	Trinidad and Tobago, Nicaragua, Sint Marteen, Germany, Austria, France, Croatia, Taiwan, Pakistan, Republic of Korea, Hong Kong, Togo
W	[0.085, 0.33]	United Kingdom, Ireland, Czec Republic, Spain, Ghana
VW	[0, 0.061]	Bahamas, Croatia, Malta, Ukraine

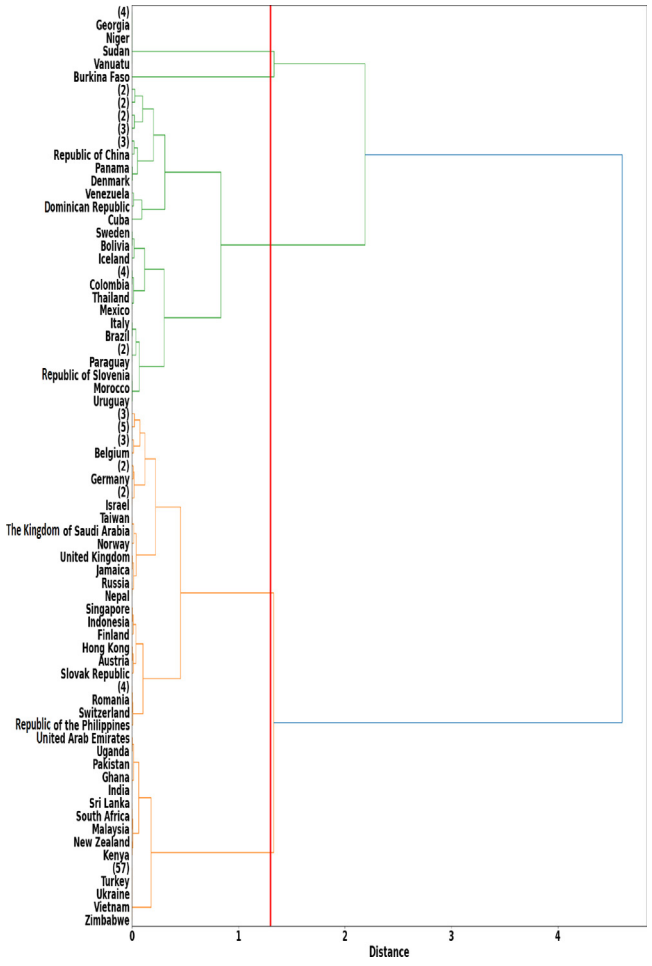

Fig. 15. Hierarchical clustering of the countries of geo-tagged tweets with respect to their Moderna RCV.

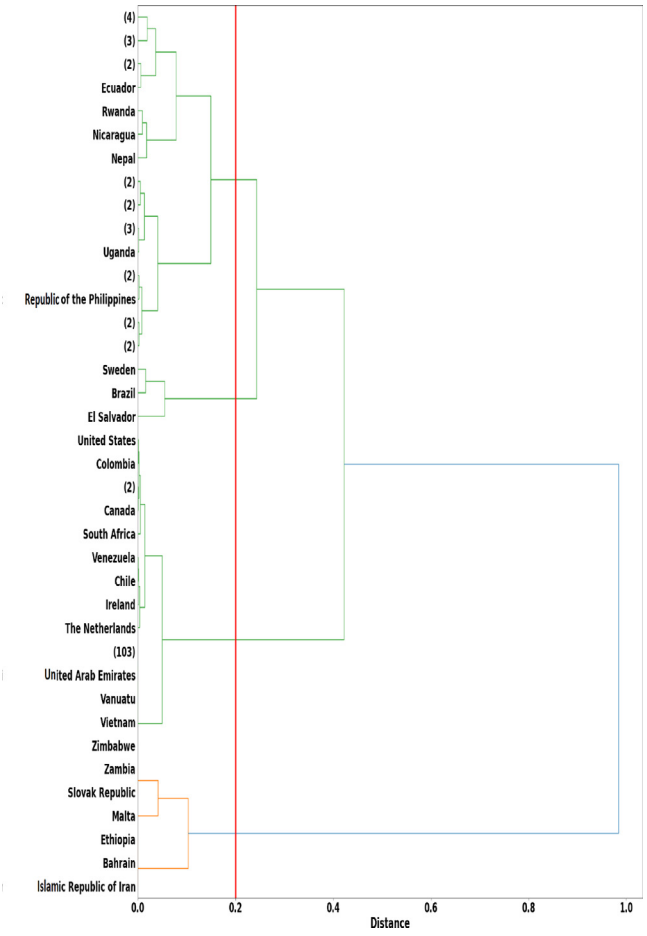

Fig. 16. Hierarchical clustering of the countries of geo-tagged tweets with respect to their AstraZeneca RCV.

not contain any Pfizer (or BioNTech) word. Our hypothesis is that the discussion was about other vaccines or about vaccination in general, without mentioning any specific name.

Fig. 15 shows the dendrogram of the countries built on the Moderna RCV values: two pairs of clusters were paired at a variance level of about 1.35, consequently, five clusters are considered. From the top of the tree, the very strong category contains 8 countries (e.g. Georgia, Niger, Sudan, Vanuatu, and Burkina Fasu), whose RCV value ranges from 1 to 2. Table 9 shows some examples of countries for each RCV category of the Moderna vaccine. Most of the South American countries (such as Mexico, Colombia, Bolivia, and Brazil) take part in the second cluster, as many South European and north European countries (e.g., Italy, Spain, Iceland, Sweden); only five countries come from the Asian continent (like Japan and China), and four from Africa (Morocco, Republic of Mozambique, Mauritius, Algeria), forming the strong category. The third cluster denotes the weak category. The 38 countries belonging to the cluster are most of the European and Asian ones. The most popular cluster contains 71 countries, denoting the very weak class. Almost all the African countries are in this cluster, as are some of the Asian ones. Furthermore, North American countries, like the Bahamas, and European ones, like Croatia, Malta, and Ukraine, mentioned the Moderna vaccine less than other countries, or even they do not mention it at all.

Fig. 16 shows the obtained clusters for the AstraZeneca RCV vaccine. The red vertical line, at a distance equal to 0.2, cuts the dendogram into four clusters. Starting from the top part of the figure, the very strong class, with a value that ranges from 0.25 to 0.33 is composed of only six countries: Bahrain, Iran, Slovak Republic, Malta, Zambia and Ethiopia, this means that at least one geo-tagged tweets out of four coming from these countries contains the vaccine name.

Table 10
Category discussion over AstraZeneca RCV values.

Category discussion	RCV range	Countries
VS	[0.25, 0.33]	Bahrain, Iran, Slovak Republic, Malta, Zambia, Ethiopia
S	[0.13, 0.19]	Sweden, El Salvador, Brazil
W	[0.021, 0.11]	Australia, Perù, Paraguay, Argentina, south Europe, Singapore, Nigeria
VW	[0, 0.018]	Ireland, Venezuela, Chile, Uruguay, Colombia, Canada, USA, India, Vietnam

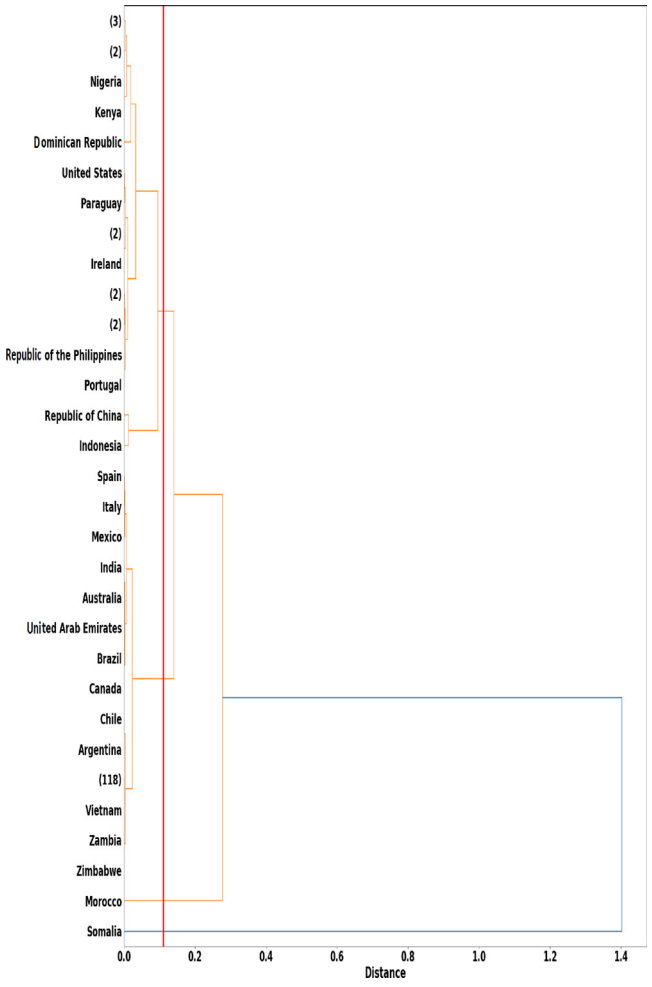

Fig. 17. Hierarchical clustering of the countries of geo-tagged tweets with respect to their Johnson&Johnson RCV.

Table 10 shows some examples of countries for each RCV category of the AstraZeneca vaccine. The second cluster is characterized by RCV values ranging from 0.13 to 0.19 and contains Sweden, El Salvador, and Brazil. The cluster corresponding to the weak class contains 28 countries with RCV between 0.021 and 0.11, among them we report Australia, Peru, Paraguay, Argentina, the south European ones, Singapore and Nigeria. The very weak cluster, with RCV values between 0.018 and 0, contains 117 countries, e.g. Ireland, Venezuela, Chile, Uruguay, Colombia, Canada, USA, India, and Vietnam.

Fig. 17 shows the dendogram of the countries built on Johnson & Johnson vaccine. The very strong class is only composed of Somalia which is considered as an outlying point due to its RCV value of 1. Table 11 shows some examples of countries for each RCV category of the Johnson & Johnson vaccine. The strong class contains only one country: Morocco, with a value of 0.2. 21 countries formed the weak cluster: Venezuela, Ecuador, Colombia, Paraguay, Belgium, Ireland, France, The Netherlands, United Kingdom, Portugal, Germany, Uganda, Kenya, Nigeria, South Africa, Malaysia, the Philippines, China, Indonesia, Dominican Republic, and the US, whose index ranges between 0.011 and 0.075. As for AstraZeneca, the geo-tagged tweets containing Johnson & Johnson vaccine are, in relative terms, fewer than the amount of Pfizer and Moderna ones, since the RCV is contained in a smaller interval. The very weak class is obtained by merging the remaining 131 countries presented in the data set, which represent all the geo-tagged tweets of locations where the vaccine is not injected or there was an absence of related posts on Twitter.

Table 11
Category discussion over Johnson & Johnson RCV values.

Category discussion	RCV range	Countries
VS	1	Somalia
S	0.2	Morocco
W	[0.011, 0.075]	Venezuela, Ecuador, Colombia, Paragua, Belgium, Ireland, France, The Netherlands, United Kingdom, Portugal, Germany, Uganda, Kenya, Nigeria, South Africa, Malaysia, the Philippines, China, Indonesia, Dominican Republic, US
VW	[0, 0.009]	Jamaica

To compare the proportion of vaccine occurrences in the considered time window, we have computed the sum of total vaccine RCV values by the continent and language pair. We have visualized this comparison by using a histogram (see Fig. 18) whose horizontal bars show the total sum of vaccine RCV values. On the vertical axis, there are a couple of labels indicating a continent (such as AF for Africa, AS for Asia, EU for Europe, NA for North America, OC for Oceania, and SA for South America) and a language (such as English, French, German, Italian, Portuguese, and Spanish) respectively. The Pfizer vaccine (represented by the blue bars) is present across almost all continents. The highest RCV values are registered in South America and in Oceania, both by German-writing tweets, with 1.225 and 1.167 respectively. Africa reaches an RCV close to 1. As regards Moderna, (represented by the orange bars) it presents a higher magnitude in Spanish, Italian, and Portuguese written posts, due to the misleading moderna term. The AstraZeneca vaccine (represented by the green bars) is more discussed in Oceania with the Spanish language (value of 0.2) and in South America with the Portuguese language (0.145). In Europe, the sum of the RCVs is near to the value 1 for every language except English: this result implies that for each geo-tagged tweet, the proper name of a vaccine is mentioned at least once. The sum increases to 1.5 for German-written tweets in South Africa and in Oceania. The Asian continent reports the smallest RCV sum values representing only the 6.95% of the total tweets as shown in Table 7.

7. Analysis of tweet texts

Text analysis activity has two purposes: the former is to define a general overview of Twitter user sentiment about the vaccination campaign roll-out; the latter is to analyze the main topics of vaccination campaigns.

As for the first aim we have employed sentiment analysis. Each tweet of our dataset has been assigned a value ranging between −1.0 and 1.0 by using *spaCy*. Relying on these values we have computed the average sentiment score for all languages by grouping the posts for each available day.

Fig. 19 shows the dynamics of sentiment scores from April 15 to September 15, 2022. Most of the lines have a magnitude around the value 0 meaning that the discussion seems to have a neutral tone over time: this might be the result of the high spreading of news including objective terms and even some comments without strongly positive or

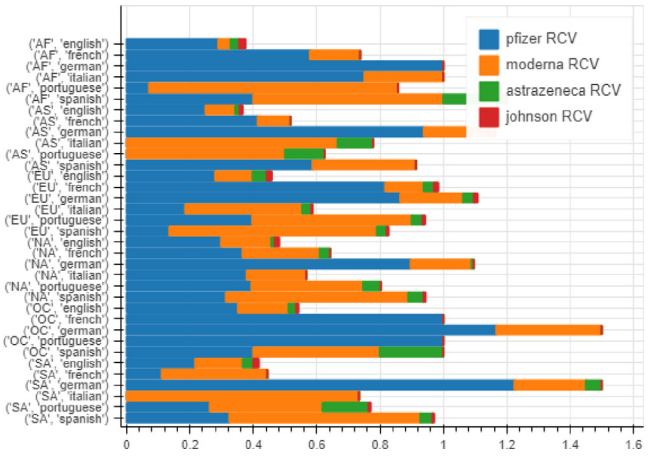

Fig. 18. Histogram of continents and languages of geo-tagged tweets with respect to their RCV value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

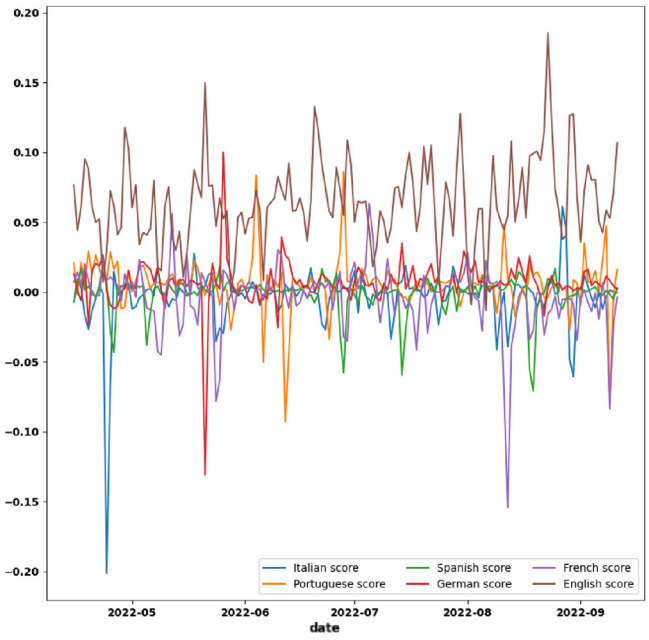

Fig. 19. Averaged sentiment scores of COVID-19 related tweets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

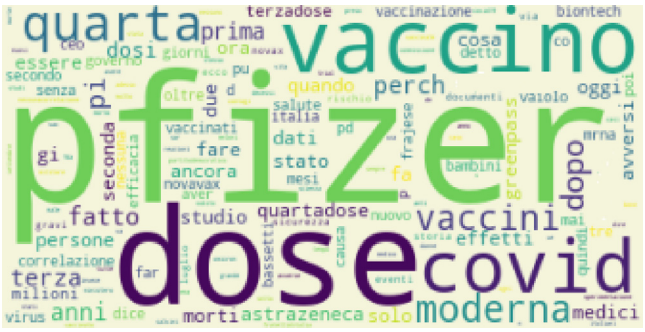

Fig. 20. Word cloud of COVID-19 related tweets in the Italian language.

Fig. 21. Word cloud of COVID-19 related tweets in English language.

negative emotions. The English average score (the brown line) has a mean of about 0.05 showing that English-written tweets have a small positive sentiment all the time. We have noted 2 peaks: one at the end of May and the other one at the end of August, with a magnitude of 0.16 and 0.18 respectively. The second one might be due to the large conversation on Moderna had on August 27, 2022. The lowest averaged sentiment score is equal to −0.20 in the Italian language (the blue line), obtained at the end of April (see Fig. 8). Except for this value, the Italian averaged score is near the zero value, with some small negative peaks in the following months. The German score (the red line) remains constant over time, with the only exception of the end of May when a negative peak of −0.13 is registered, and the first days of June, when there is a drastic trend reversal of magnitude 0.10. The French score (the purple line) is lower than 0 on average with a few peaks. The worst value is obtained in mid-August, it might be the result of the Astrazeneca count peak of August 17, 2022. The Spanish score trend (the green line) seems to be similar to the French one: a lot of negative peaks around −0.05 are widespread across the time window, and no significant averaged positive scores are recorded. The Portuguese sentiment (the yellow line) is characterized by several trend changes one in the first half of June, from 0.10 to −0.05, and another one in the second half of the same month, from −0.03 to a 0.09 in less than a week. The final part of the line shows some small positive peaks and a negative one in the mid of September.

As regards topic analysis we have employed word cloud and topic modeling techniques which rely on the fact word occurrences are supposed to be the significant drivers of vaccine contextualization.

We have written code in python to visually represent the most commonly used words within the language-specific tweets using the *wordcloud* package [56]. We have chosen to display the most frequent 150 terms for each language: the more frequent a term, the bigger its font size. Table 12 shows the main considered terms that appear in the various language-specific word clouds figures and topic modeling tables with the relative English translation used in the following paragraphs.

The Pfizer and dose terms are the most frequent words in the Italian word cloud (see in Fig. 20) as well as covid, moderna, fourth, and vaccine. Political terms have been detected, such as government, the ex-Italian Prime Minister, and the former Minister of Health respectively. The efficacy on people was also discussed, since effects, adverse, and children are reported.

Fig. 21 shows that the vaccine, covid, pfizer, moderna, and vaccination terms are the most common among English Twitter users. Furthermore, population health is discussed through the following words: people, death, health, children, pregnant, and age. Several aspects regarding vaccines have been mentioned, e.g., shot, dose, booster, effect. AstraZeneca and Johnson & Johnson present a very low frequency, as the terms have been reported in a pretty small font size. The fda term is related to the Food and Drug Administration American government agency.

The Spanish word cloud (see Fig. 22) reports Pfizer, Moderna, and vaccine, as the most common words. The Moderna term may have different meanings according to the context and its frequency seems to be higher than Pfizer. A possible explanation is that moderna could have

Table 12
Terms Translation in the English Language.

English	French	German	Italian	Portuguese	Spanish
adverse	effets		avversi		
arms			armi		
biological			biologiche		
biological arms			armi biologiche		
cardiac				cardíaca	
children		kinder	bambini		nios
correlation			correlazioni		
diseases				doenças	
document	document		documento	documento	documento
dose			dose		
doses					dosis
effects	effet		effetti		
effectiveness			efficacia		
efficiency		wirksamkeit			
first			prima		
fourth			quarta	quarta	
friendship				amizade	
future					futuro
government			governo		
gurbage			spazzatura		basura
heart				corao	
history					historia
laughing				rindo	
life					vida
maternity			gravidanza		
millions					milliones
minister			ministero		
myocardium				miocárdio	
modern					moderna
pay					pago
person					pessoa
people				pessoas	
pregnant			incinte		
president					presidente
psicological				psicologico	
requests			richieste		
resonances				ressonâcias	
health			salute		
salary					sueldo
seasonal			stagionale		
secondary	secondaires				
third			terza		
impostor	truqu				
vaccine		impfstoff	vaccino	vacina	vacuna
vaccination		impfung			
witness	témoin				
women	femmes		donne		mujeres

been used as an adjective of the Spanish translation of history since some peaks in the respective time series (see Fig. 10) denoted exactly tweets talking about modern and contemporary history. AstraZeneca shows a higher frequency than in the previous Figures; this may be the consequence of an almost absent Astrazeneca vaccination in English-speaking countries with respect to Spanish ones. Mexico seems to be the most mentioned country because of the presence of the word mxico, whose e letter has been dropped in the pre-processing step because of its accents such as *é*, *è*.

In the Portuguese word cloud (see Fig. 23), the Pfizer, Moderna and vaccine terms are the top frequent three words. The second most mentioned terms are covid, AstraZeneca and fourth; dose and person follow the list. The document word may reveal that people tended to inform themselves through a government or medical document. Moreover, the Portuguese-speaking people showed empathy towards other people since the friendship and psychological, appear in the Figure.

In the French word cloud (see Fig. 24) Pfizer, Moderna, vaccine, covid, and vaccination are the most common terms. After those, the name of the French president and of the EU Commission appear. Furthermore, the document word might relate to the documentation of some COVID-19 related news that was discussed by French-speaking users. In addition, medical terms like dose, injection, and effect have been reported.

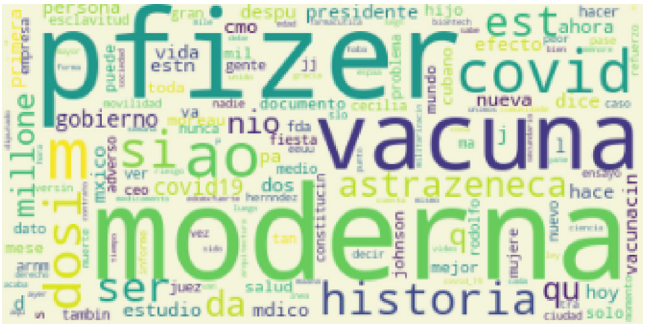

Fig. 22. Word cloud of COVID-19 related tweets in Spanish language.

In the German word cloud (see Fig. 25) the words that have been more frequently written are Pfizer, Biontech and Moderna. The second most frequent words are vaccine and vaccination. The corona term, an abbreviation of coronavirus, appears only in this word cloud, as the Paxlovid term that is the name of the anti-COVID drug, released at the end of January 2022.

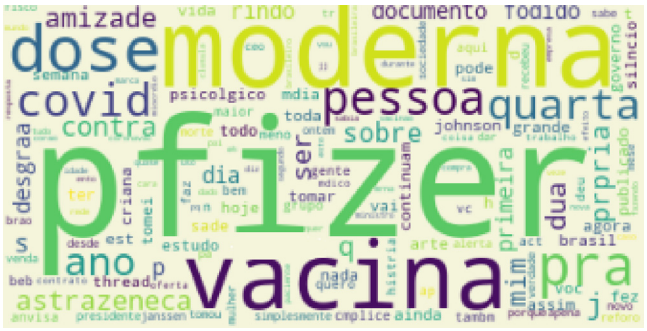

Fig. 23. Word cloud of COVID-19 related tweets in Portuguese language.

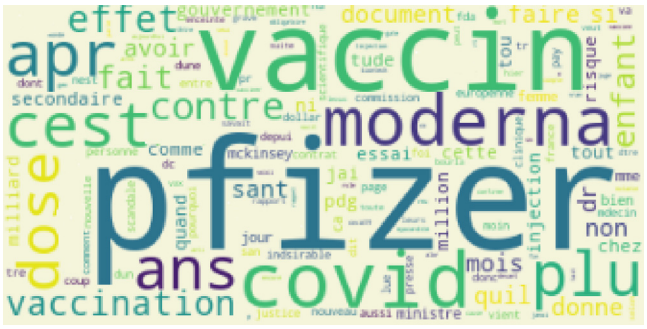

Fig. 24. Word cloud of COVID-19 related tweets in French language.

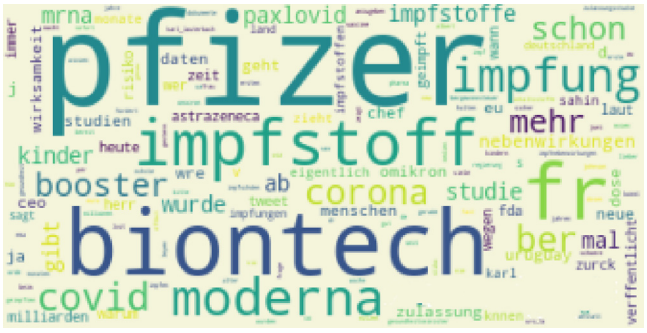

Fig. 25. Word cloud of COVID-19 related tweets in German language.

Topic modeling has been applied to the outcome of the tweet analysis based on TF-IDF and LDA models, mentioned in Section 3.4. This technique requires the specification of a target number of topics: we have chosen a value between 2 and 10, with the aim of having a broad spectrum of main themes. In the following tables, the 10 words with the highest weight for each topic with their coherence score are reported. Furthermore, we have added a possible interpretation of the most important themes. The accented characters that have been removed in the pre-processing step are shown in round brackets for a better understanding of the topics.

The Italian topic modeling in Table 13 reports 9 major discussions with a coherence of 0.4303, meaning that this result is adaptable to tweets written in Italian for about 43%. The first topic includes words like seasonal, garbage, and serious, inferring the alleged correlation between seasonal jobs and the increasing requests for vaccines. The second topic contains tokens such as women, pregnant, and zoster, showing interest in the supposed larger fragility of pregnant women and in the herpes zoster subject. The third topic lists terms like effectiveness, biology, and arms that may point out the presence of biological arms in vaccines. Italian-writing people also discussed the correlation that could be present between booster doses and monkeypox development. Apart from that, several words like after, adverse, and doses, are mentioned in the other topics, indicating concerns about vaccination campaigns and side effects.

As for the English tweets, 8 topics have been identified with a coherence of 0.3395 (see Table 14), its value, lower than the Italian one, is a consequence of the larger number of English tweets with respect to the other languages. The first topic includes words like documents, effective, and mandatory, they could show interest in the effectiveness of vaccines. The second topic contains tokens like discharged, cult, and incurable, they may concern the vaccine roll-out. The third topic contains tokens like pregnant, women, and babies, conveying concern about the consequences of vaccines. Apart from that, several words like vaccineinjuries, vaccinedeaths, and hospital, are mentioned in the other topics, indicating concerns about vaccination effects and their trustworthiness.

As regards the Spanish tweets, 8 themes have been detected with a coherence of 0.4177 (see Table 15). The first topic contains tokens like pregnant, women, and monkeypox, they could show interest in the consequences of vaccines on pregnant women. The fifth topic includes terms like smallpox, future, and world, inferring some concerns about the spread of the disease in the future. The sixth topic lists words such as millions, doses, and vaccine, they seem to show interest in how many doses are necessary to provide for the whole population. The seventh topic contains tokens like pay, salary, and life, conveying concerns about life from an economic point of view.

As for Portuguese tweets, they can be grouped into 10 topics, with a coherence of 0.3586 (see Table 16). The first topic contains tokens like heart, myocardium, and resonances inferring concerns about the consequences of vaccination. The third topic includes words such as friendship, people, and laughing showing the necessity for people to be together. The fourth topic lists terms like cardiac, death, and disease, these words may be related to heart-related discussion.

Regarding French tweets, 9 topics have been identified, with a coherence of 0.3848 (see Table 17). The third topic includes words such as adverse, secondary, reaction, and women, showing concerns about the side effects of vaccination.

As regards German tweets, only 2 topics have been detected, with a coherence of 0.4757 (see Table 18). The first topic lists terms like booster, effects, covid, showing interest in the effects of the booster doses of vaccines. The second topic contains tokens such as efficiency, research, and children conveying concerns about the relationship between vaccine and children.

8. Discussion and conclusions

The spread of the COVID-19 pandemic has dominated online conversations on social media. The content posted by users can be used to perform a rich spectrum of analysis, from counting tweets or keywords to content analysis, by working with textual data. The majority of the existing studies that has investigated the relationship between social media and vaccination have focused on specific areas or languages. Our work is one of the few that address the topic of COVID-19 vaccination relying on multi-language posts. It has considered tweets all over the world written in different languages to identify the main vaccine brand discussed among people on social media. Furthermore, we have tried to understand public opinion by interpreting the topics highlighted from the terms in the tweets to detect how the vaccine campaign is accepted by people and therefore to help authorities make decisions to promote their campaign.

In our study, after the pre-processing phase, we have analyzed 8,343,490 tweets. The answers to our four questions related to the COVID-19 vaccination campaign are as follows:

Table 13
Topic modeling of COVID-19 related tweets in the Italian language.

Topic	Words
1	spazzatura, gravi, pfizer, stagionale, eventi, rischio, boom, vaccino, definisce, richieste
2	donne, incinte, cavolo, vaccinazione, documenti, zoster, gravidanza, pfizer, topi, vaccino
3	pfizer, moderna, armi, efficacia, biologiche, bastava, ammette, signora
4	morivano, moderna, game, falso, vaccinati, dose, california, sapeva, pfizer, ufficialmente
5	novavax, correlazione, nessuna, greenpass, quartadose, terzadose, monkeypox, moderna, vaccino
6	ministero, difesa, salute, pfizer, quando, russo, eventiavversi, ulteriori, colpito
7	dose, quarta, dopo, fatto, pfizer, terza, vaccino, seconda, covid, prima
8	vaccini, covid, vaccino, pfizer, pagine, tutela, efficace, mila, cavie, dopo
9	davos, milioni, pfizer, dopo, covid, dose, vaccini, vaccino, vacc, moderna
Coherence	0.4303

Table 14
Topic modeling of COVID-19 related tweets in the English language.

Topic	Words
1	pfizer, documents, effective, people, vaccine, vaccines, knew, covid, never, mandatory
2	discharged, cult, king, johnson, davos, bourla, albert, incurable, vikki, realise
3	pregnant, women, babies, sued, director, confirms, head, pfizer, told, facebook
4	vaccineinjuries, vaccinedeaths, pfizerdocuments, canadians, genocide, criminals, brought, scope, pots, reviewed
5	injected, lost, pain, house, lawyers, reveal, hunt, ramsay, documents, chest
6	covid, vaccine, teenagers, children, vaccination, moderna, says, without, june, vaccines
7	quoting, holland, guillain, barr(è), holy, explains, chip, hearts, sads, ingestible
8	days, recently, myocarditis, heart, mrna, hospital, risk, vaccine, covid, syndrome
Coherence	0.3395

Table 15
Topic modeling of COVID-19 related tweets in the Spanish language.

Topic	Words
1	embarazadas, naciones, mujeres, documentos, astrazeneca, obama, intelectuales, vacuna, pfizer, chimpanc(è)
2	bullrich, pudo, basura, fallo, explicar, nadie, arquitecto, tirar, trabajadores, quiere
3	cubanos, m(é)dicos, cuba, covid, pfizer, pidi(ó), esclavos, personas, esclavitud, probar
4	moderna, movilidad, comunidades, unimos, edomxfuerte, exponentes, m(á)ximos, recordamos, vida, mono
5	moderna, presindente, denucian, historia, mono, futuro, mundo, nueva, viruela, tratados
6	pfizer, dosis, vacuna, covid, vacunas, nios, astrazeneca, milliones, moderna, casos
7	vida, larga, temporadas, programas, moderna, vicepresidente, pago, sueldo, medio, triste
8	villanueva, arquitectura, carlos, constitucin, demand, gladysmarn, preguntaron, respondi, desallorato, moderna
Coherence	0.4177

Table 16
Topic modeling of COVID-19 related tweets in the Portuguese language.

Topic	Words
1	simplesmente, ressonn(â)cias, queixas, pacientes, mioc(á)rdio, corao, alerta, quero, robert, malone
2	pfizer, thread, publicados, continuam, sil(ê)ncio, c(ú)mplice, m(é)dia, documentos, grande, primeira
3	pr(ó)pria, rindo, desgra(ç)a, psicol(ó)gico, duas, amizade, pessoas, fodido, moderna, fonte
4	johnson, explode, s(ú)bitas, irish, irlands, light, capa, card(í)aca, doen(ç)as, mortes
5	dose, quarta, covid, vacina, anos, contra, tomar, pfizer, sade, doses
6	p(á)ginas, publicar, forou, mark, pittman, interno, federal, juiz, trabalho, aqui
7	venda, bourla, brasil, nada, menos, tomou, cl(á)usulas, albert, vacinas, absurdas
8	astrazeneca, quinto, ainda, recomenda, morreram, moderna, refor(ç)o, saber, pfizer, pagar
9	anos, arte, clinicas, moderna, aconselham, particulares, look, davos, documenta(çã)o, jovens
10	debutava, more, dream, sucedidos, single, grupo, todos, hoje, anos, assim
Coherence	0.3586

Table 17
Topic modeling of COVID-19 related tweets in the French language.

Topic	Words
1	courrier, condamn, syst(è)me, tour, vient, nouvelle, lagence, sour, devrait, honn(ê)te
2	ctait, raoult, doses, enfants, pfizer, r(é)sultats, mckinsey, truqu, t(é)moin
3	effets, ind(é)siderables, vaccin, femmes, documents, secondaires, pages, enceintes, pfizer, montrent
4	secrets, diffuser, oblige, tombent, masques, justice, nont, hier, documents, vaccin
5	m(é)decine, donne, page, vaccinale, fran(ç)aise, rappelez, strat(é)gie, lors, heures, mari
6	devant, suspendus, soignants, silvano, trotta, surprise, censure, lars, motion, covid
7	fois, paxlovid, bourla, vaccin, millions, gates, suite, covid, france, command
8	veux, mensognes, afin, larnaque, suivre, covidiste, sensibilis(é)s, puissions, suspendue, demand
9	quotidien, schwab, klaus, lansm, cliniques, d(é)cryptes, reconnu, aujourdhui, guillain, utilisation
Coherence	0.3848

Table 18
Topic modeling of COVID-19 related tweets in the German language.

Topic	Words
1	pfizer, covaxin, moderna, biontech, covid, booster, effects, order, alphabetical
2	biontech, impfstoff, impfung, moderna, kinder, wurde, studie, wirksamkeit, daten, gibt
Coherence	0.4757

- RQ1 What have we learned from COVID-19 vaccination in general? To reply to this question, we have considered our geo-tagged tweets and text analysis results. Tweet trends over time seem to follow similar patterns as we have noticed in time series charts. These charts illustrate the values of the RCV index that has been essential in identifying peaks attributed to COVID-19 vaccination-related incidents. News related to the success or failure of vaccination is immediately reflected on social media as shown in several tweets.
- RQ2 What is the impact of language on tweet-based research about COVID-19 vaccines? Language seems not to be relevant in the discussion about COVID-19 vaccines apart from when there are confusing terms, e.g. Moderna in Spanish. However, a multi-language analysis could facilitate the collation of diverse opinions and perspectives from varied cultural contexts regarding a pandemic that has drastically impacted lives across the globe.
- RQ3 Which brands of COVID-19 vaccine have been the most debated in different countries and languages? The most debated vaccine all over the world is Pfizer, according to Fig. 18 and time series results.
- RQ4 What have been the main language-specific topics of discussion regarding COVID-19 vaccines? Concerns among individuals have been observed pertaining to the impact of vaccines on pregnant women, children, and heart conditions. Generally speaking, skepticism regarding the efficacy and potential risks associated with vaccines has intensified over the months.

In conclusion, our analysis has highlighted how the COVID-19 vaccination campaign has been affected by news regarding vaccine side effects. Our outcomes are in line with other studies that refer to the first months of the vaccination campaigns. The general sentiment about COVID-19 vaccination campaigns remained constant and mainly neutral, proving that the hesitancy rate and the opinions during the booster dose roll-out are somewhat unchanged over the analyzed timestamp. The possible presence of controversial theories and negative statements regarding vaccines efficacy seems to be balanced with positive emotions and good news instantly reported on Twitter. Hence, our analysis showed that worldwide people continue to get vaccinated, even though it is no longer mandatory.

Further studies could investigate the opinions about COVID-19 vaccines in the coming months and years. Furthermore, our solution can be extended to identify misleading information given by the language.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] A. Aleksandric, H.I. Anderson, S. Melcher, S. Nilizadeh, G.M. Wilson, Spanish Facebook posts as an indicator of COVID-19 vaccine hesitancy in Texas, Vaccines 10 (2022) http://dx.doi.org/10.3390/vaccines10101713.

[2] K. Lanyi, R. Green, D. Craig, M. C., COVID-19 vaccine hesitancy: Analysing Twitter to identify barriers to vaccination in a Low Uptake Region of the UK, Front. Digit. Health 24 (2022) http://dx.doi.org/10.3389/fdgth.2021.804855.

[3] D. Wawrzuta, J. Klejdysz, M. Jaworski, J. Gotlib, M. Panczyk, Attitudes toward COVID-19 vaccination on social media: A cross-platform analysis, Vaccines 10 (8) (2022) http://dx.doi.org/10.3390/vaccines10081190.

[4] M. Mahdikhani, Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of COVID-19 pandemic, Int. J. Inf. Manag. Data Insights 2 (1) (2022) 100053, http://dx.doi.org/10.1016/j.jjimei.2021.100053.

[5] İ. Aygün, B. Kaya, M. Kaya, Aspect based twitter sentiment analysis on vaccination and vaccine types in COVID-19 pandemic with deep learning, IEEE J. Biomed. Health Inform. 26 (5) (2022) 2360–2369, http://dx.doi.org/10.1109/JBHI.2021.3133103.

[6] J. Cornelius, T. Ellendorff, L. Furrer, F. Rinaldi, COVID-19 Twitter monitor: Aggregating and visualizing COVID-19 related trends in social media, in: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1–10.

[7] L. Li, A. Aldosery, F. Vitiugin, N. Nathan, D. Novillo-Ortiz, C. Castillo, P. Kostkova, The response of governments and public health agencies to COVID-19 pandemics on social media: A multi-country analysis of Twitter discourse, Front. Public Health 9 (2021) http://dx.doi.org/10.3389/fpubh.2021.716333.

[8] L. Huangfu, Y. Mo, P. Zhang, H.S. Zeng DD, COVID-19 vaccine tweets after vaccine rollout: Sentiment-based topic modeling, J. Med. Internet Res. 24 (2022) http://dx.doi.org/10.2196/31726.

[9] H. Jang, E. Rempel, I. Roe, P. Adu, G. Carenini, N.Z. Janjua, Tracking public attitudes toward COVID-19 vaccination on tweets in Canada: Using aspect-based sentiment analysis, J. Med. Internet Res. 24 (3) (2022) e35016, http://dx.doi.org/10.2196/35016.

[10] A.A. El-Latief, Analysis of US COVID-19 Twitter data social interest and topic changes, in: 2021 17th International Computer Engineering Conference, ICENCO, 2021, pp. 14–17, http://dx.doi.org/10.1109/ICENCO49852.2021.9698933.

[11] B. Ogbuokiri, A. Ahmadi, N.L. Bragazzi, Z. Movahedi Nia, B. Mellado, J. Wu, J. Orbinski, A. Asgary, J. Kong, Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts.

[12] L. Stracqualursi, P. Agati, COVID-19 vaccines in Italian public opinion: Identifying key issues using Twitter and natural language processing, PLoS One 17 (11) (2022) 1–18, http://dx.doi.org/10.1371/journal.pone.0277394.

[13] P. Biancovilli, L. Makszin, C. Jurberg, Misinformation on social networks during the novel coronavirus pandemic: A quali-quantitative case study of Brazil, BMC Public Health 21 (2021) 1–10.

[14] S. Andreadis, G. Antzoulatos, T. Mavropoulos, P. Giannakeris, G. Tzionis, N. Pantelidis, K. Ioannidis, A. Karakostas, I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris, A social media analytics platform visualising the spread of COVID-19 in Italy via exploitation of automatically geotagged tweets, Online Soc. Netw. Media 23 (2021) 100134, http://dx.doi.org/10.1016/j.osnem.2021.100134.

[15] A. Umair, E. Masciari, G. Madeo, M.H. Ullah, Sentimental analysis of COVID-19 vaccine tweets using bert+nbsvm, in: I. Koprinska, P. Mignone, R. Guidotti, S. Jaroszewicz, H. Fröning, F. Gullo, P.M. Ferreira, D. Roqueiro, G. Ceddia, S. Nowaczyk, J.a. Gama, R. Ribeiro, R. Gavaldà, E. Masciari, Z. Ras, E. Ritacco, F. Naretto, A. Theissler, P. Biecek, W. Verbeke, G. Schiele, F. Pernkopf, M. Blott, I. Bordino, I.L. Danesi, G. Ponti, L. Severini, A. Appice, G. Andresini, I. Medeiros, G. Graça, L. Cooper, N. Ghazaleh, J. Richiardi, D. Saldana, K. Sechidis, A. Canakoglu, S. Pido, P. Pinoli, A. Bifet, S. Pashami (Eds.), Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer Nature Switzerland, Cham, 2023, pp. 238–247.

[16] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, P. Cotae, COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset, Expert Syst. Appl. 212 (2023) 118715, http://dx.doi.org/10.1016/j.eswa.2022.118715, URL https://www.sciencedirect.com/science/article/pii/S0957417422017407.

[17] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, S. Sharif, An analysis of COVID-19 vaccine sentiments and opinions on Twitter, Int. J. Infect. Dis. 108 (2021) 256–262, http://dx.doi.org/10.1016/j.ijid.2021.05.059.

[18] N. Puri, E.A. Coomes, H. Haghbayan, K. Gunaratne, Social media and vaccine hesitancy: New updates for the era of COVID-19 and globalized infectious diseases, Hum. Vaccines Immunother. 16 (11) (2020) 2586–2593, http://dx.doi.org/10.1080/21645515.2020.1780846.

[19] W.-Y.S. Chou, A. Budenz, Considering emotion in COVID-19 vaccine communication: Addressing vaccine hesitancy and fostering vaccine confidence, Health Commun. 35 (14) (2020) 1718–1722, http://dx.doi.org/10.1080/10410236.2020.1838096.

[20] K.L. Sussman, L. Bouchacourt, L.F. Bright, G.B. Wilcox, M. Mackert, A.S. Norwood, B.S. Allport Altillo, COVID-19 topics and emotional frames in vaccine hesitation: A social media text and sentiment analysis, Dig. Health 9 (2023) http://dx.doi.org/10.1177/20552076231158308.

[21] A. Aleksandric, M.J. Obasanya, S. Melcher, S. Nilizadeh, G.M. Wilson, Your tweets matter: How social media sentiments associate with COVID-19 vaccination rates in the US, Online J. Public Health Inform. 14 (1) (2022) http://dx.doi.org/10.5210/ojphi.v14i1.12419.

[22] S. Mishra, A. Verma, K. Meena, R. Kaushal, Public reactions towards COVID-19 vaccination through twitter before and after second wave in India, Soc. Netw. Anal. Min. 12 (1) (2022) 57, http://dx.doi.org/10.1007/s13278-022-00885-w.

[23] A. Ljaji'c, N. Prodanovi'c, D. Medvecki, B. Bašaragin, J. Mitrovi'c, Uncovering the reasons behind COVID-19 vaccine hesitancy in Serbia: Sentiment-based topic modeling, J. Med. Internet Res. 24 (11) (2022) http://dx.doi.org/10.2196/42261.

[24] S.W.H. Kwok, S.K. Vadde, G. Wang, Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis, J. Med. Internet Res. 23 (5) (2021) e26953, http://dx.doi.org/10.2196/26953.

[25] S. Ahmed, M.E. Rasul, J. Cho, Social media news use induces COVID-19 vaccine hesitancy through skepticism regarding its efficacy: A longitudinal study from the united states, Front. Psychol. 13 (2022) http://dx.doi.org/10.3389/fpsyg.2022.900386, URL https://www.frontiersin.org/articles/10.3389/fpsyg.2022.900386.

[26] B. Ogbuokiri, A. Ahmadi, Z.M. Nia, B. Mellado, J. Wu, J. Orbinski, A. Asgary, J. Kong, Vaccine hesitancy hotspots in Africa: An insight from geotagged Twitter posts, IEEE Trans. Comput. Soc. Syst. (2023) 1–14, http://dx.doi.org/10.1109/TCSS.2023.3236368.

[27] R.Y. nez Martínez, G. Blanco, A. Lourenço, Spanish corpora of tweets about COVID-19 vaccination for automatic stance detection, Inf. Process. Manage. 60 (3) (2023) http://dx.doi.org/10.1016/j.ipm.2023.103294, URL https://www.sciencedirect.com/science/article/pii/S0306457323000316.

[28] G. Lindelöf, T. Aledavood, B. Keller, Dynamics of negative discourse toward COVID-19 vaccines: A topic modeling study and an annotated dataset of Twitter posts, J. Med. Internet Res. (2023) http://dx.doi.org/10.2196/41319.

[29] R.P. Alsudias L, Social media monitoring of the COVID-19 pandemic and influenza epidemic with adaptation for informal language in Arabic Twitter data: Qualitative study, JMIR Med. Inform. 17 (2021).

[30] D. Gori, C. Reno, D. Remondini, F. Durazzi, M.P. Fantini, Are we ready for the arrival of the new COVID-19 vaccinations? Great promises and unknown challenges still to come, Vaccines 9 (2) (2021) 173, http://dx.doi.org/10.3390/vaccines9020173.

[31] J. Griffith, H. Marani, H. Monkman, COVID-19 vaccine hesitancy in Canada: Content analysis of tweetsx using the theoretical domains framework, J. Med. Internet Res. 23 (4) (2021) e26874, http://dx.doi.org/10.2196/26874.

[32] S. Fazel, L. Zhang, B. Javid, et al., Harnessing Twitter data to survey public attention and attitudes towards COVID-19 vaccines in the UK, Sci. Rep. 11 (1) (2021) http://dx.doi.org/10.1038/s41598-021-02710-4.

[33] R. Marcec, R. Likic, Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and moderna COVID-19 vaccines, Postgrad. Med. J. 98 (1161) (2022) 544–550, http://dx.doi.org/10.1136/postgradmedj-2021-140685, arXiv:https://pmj.bmj.com/content/98/1161/544.full.pdf.

[34] N. Chen, X. Chen, J. Pang, A multilingual dataset of COVID-19 vaccination attitudes on Twitter, Data Brief 44 (2022) 108503, http://dx.doi.org/10.1016/j.dib.2022.108503.

[35] A. Sepúlveda, C. Periñán Pascual, A. Muñoz, R. Martí nez España, E. Hernández-Orallo, J.M. Cecilia, COVIDSensing: social sensing strategy for the management of the COVID-19 crisis, Electronics 10 (24) (2021) 3157, http://dx.doi.org/10.3390/electronics10243157.

[36] J.M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, E. Artemova, E. Tutubalina, G. Chowell, A large-scale COVID-19 Twitter chatter dataset for open scientific research—An international collaboration, Epidemiologia 2 (3) (2021) 315–324, http://dx.doi.org/10.3390/epidemiologia2030024.

[37] S.C. Guntuku, A.M. Buttenheim, G. Sherman, R.M. Merchant, Twitter discourse reveals geographical and temporal variation in concerns about COVID-19 vaccines in the United States, Vaccine 39 (30) (2021) 4034–4038, http://dx.doi.org/10.1016/j.vaccine.2021.06.014.

[38] TSAPI, Stream Tweets in real-time, https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time.

[39] tweepy, Tweepy Documentation, Available online, https://docs.tweepy.org/en/stable/. (Accessed 19 February 2023).

[40] A. Bari, M. Heymann, R. Cohen, R. Zhao, L. Szabo, S. Vasandani, A. Khubchandani, M. DiLorenzo, M. Coffee, Exploring coronavirus disease 2019 vaccine hesitancy on Twitter using sentiment analysis and natural language processing algorithms, Clin. Infect. Dis. 74 (2022) e4–e9, http://dx.doi.org/10.1093/cid/ciac141.

[41] L.B. Shyamasundar, P. Jhansi Rani, A multiple-layer machine learning architecture for improved accuracy in sentiment analysis, Comput. J. 63 (1) (2020) 395–409, http://dx.doi.org/10.1093/comjnl/bxz038.

[42] ISO, ISO 3166 Country Codes, https://www.iso.org/iso-3166-country-codes.html.

[43] M. Kazijevs, F.A. Akyelken, M.D. Samad, Mining social media data to predict COVID-19 case counts, in: 2022 IEEE 10th International Conference on Healthcare Informatics, ICHI, 2022, pp. 104–111, http://dx.doi.org/10.1109/ICHI54592.2022.00027.

[44] Y. Emre Palavar, Z. Çelk, A. Albayrak, E. Özçelk, M. Erdl, Analysis of the COVID-19 process in terms of health managers, in: 2022 2nd International Conference on Computing and Machine Intelligence, ICMI, 2022, pp. 1–5, http://dx.doi.org/10.1109/ICMI55296.2022.9873660.

[45] T. Hu, B. She, L. Duan, H. Yue, J. Clunis, A systematic spatial and temporal sentiment analysis on geo-tweets, IEEE Access 8 (2020) 8658–8667, http://dx.doi.org/10.1109/ACCESS.2019.2961100.

[46] Q. Khan, H.N. Chua, Comparing topic modeling techniques for identifying informative and uninformative content: A case study on COVID-19 tweets, in: 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET, 2021, pp. 1–6, http://dx.doi.org/10.1109/IICAIET51634.2021.9573878.

[47] S. Perera, I. Perera, S. Ahangama, Exploring Twitter messages during the COVID-19 pandemic in Sri Lanka: Topic modelling and emotion analysis, in: 2022 2nd International Conference on Advanced Research in Computing, ICARC, 2022, pp. 148–153, http://dx.doi.org/10.1109/ICARC54489.2022.9753866.

[48] Spacy, Industrial-Strength Natural Language Processing in Python, https://spacy.io/.

[49] Folium, Python Library, https://pypi.org/project/folium/.

[50] A. Doroshenko, Analysis of the distribution of COVID-19 in Italy using clustering algorithms, in: 2020 IEEE Third International Conference on Data Stream Mining & Processing, DSMP, 2020, pp. 325–328, http://dx.doi.org/10.1109/DSMP47368.2020.9204202.

[51] W.B. Zulfikar, A. Wahana, R.D. Sukma, D.R. Ramdania, D.S. Maylawati, Game popularity level during COVID-19 pandemic using agglomerative hierarchical clustering, in: 2022 10th International Conference on Cyber and IT Service Management, CITSM, 2022, pp. 1–5, http://dx.doi.org/10.1109/CITSM56380.2022.9936040.

[52] S. Kumar, Use of cluster analysis to monitor novel coronavirus-19 infections in maharashtra, India, Indian J. Med. Sci. 72 (2) (2020) 44–48, http://dx.doi.org/10.25259/IJMS_68_2020.

[53] F. Durazzi, M. Müller, S. M., D. Remondini, Clusters of science and health related Twitter users become more isolated during the COVID-19 pandemic, Sci. Rep. 11 (2021) 75–174, http://dx.doi.org/10.1038/s41598-021-99301-0.

[54] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Amer. Statist. Assoc. 58 (301) (1963) 236–244, http://dx.doi.org/10.2307/2282967.

[55] C.K. Reddy, B. Vinzamuri, A survey of partitional and hierarchical clustering algorithms, in: Data Clustering: Algorithms and Applications, 2018.

[56] A. Mueller, WordCloud for Python documentation, 2020, Available online, https://amueller.github.io/word_cloud/. (Accessed 19 February 2023).