Exploring Pandemics Events on Twitter by Using Sentiment Analysis and Topic Modelling

Zhikang Qin 1 and Elisabetta Ronchieri 1,2,*

1 Department of Statistical Sciences, University of Bologna, 40126 Bologna, Italy
2 INFN National Institute for Nuclear Physics CNAF, 40126 Bologna, Italy
* Correspondence: elisabetta.ronchieri@cnaf.infn.it; Tel.: +39-0512095072

Abstract: At the end of 2019, while the world was being hit by the COVID-19 virus and, consequently, was living a global health crisis, many other pandemics were putting humankind in danger. The role of social media is of paramount importance in these kinds of contexts because they help health systems to cope with emergencies by contributing to conducting some activities, such as the identification of public concerns, the detection of infections’ symptoms, and the traceability of the virus diffusion. In this paper, we have analysed comments on events related to cholera, Ebola, HIV/AIDS, influenza, malaria, Spanish influenza, swine flu, tuberculosis, typhus, yellow fever, and Zika, collecting 369,472 tweets from 3 March to 15 September 2022. Our analysis has started with the collection of comments composed of unstructured texts on which we have applied natural language processing solutions. Following, we have employed topic modelling and sentiment analysis techniques to obtain a collection of people’s concerns and attitudes towards these pandemics. According to our findings, people’s discussions were mostly about malaria, influenza, and tuberculosis, and the focus was on the diseases themselves. As regards emotions, the most popular were fear, trust, and disgust, where trust is mainly regarding HIV/AIDS tweets.

Keywords: epidemics; Twitter; natural language processing; topic modelling; sentiment analysis; ARI; cholera; Ebola; malaria; Zika

1. Introduction

Pandemics represent a threat to human survival. Infectious diseases are responsible for many deaths [1] and inflict a burden on public health systems [2]. Recently, COVID-19 has ravaged the globe, becoming a hot spot for research. However, COVID-19 is just one of the pandemics that causes suffering on our planet. Social media represents a valid instrument to understand public perceptions during a pandemic, providing some guidelines to governments and medical organisations.

In this work, we have collected tweets—written from 3 March to 15 September 2022—about 11 pandemics, including cholera, Ebola, HIV/AIDS, influenza, malaria, Spanish influenza, swine flu, tuberculosis, typhus, yellow fever, and Zika (all of them are detailed in Appendix A). The considered tweets have been extracted from Twitter, one of the most famous mobile microblogging and social networking services in the world. As proved in the previous literature, social media has become important for public health surveillance and monitoring [3,4]; therefore, our study has exploited tweets to reveal public opinions in the course of pandemic events with the aim of identifying the key factors of public interest to limit the spread of the disease.

Working mainly with unstructured texts, we have applied natural language processing (NLP) techniques [5,6] to analyse epidemic-related messages on Twitter. By using several machine learning techniques, we aim to answer the following research questions: RQ1 Which pandemics have more discussion on social media? What is the trend of these discussions over time?
RQ2 What are people’s concerns related to these pandemics?
RQ3 What are people’s attitudes or emotions to these epidemics?
RQ4 How can the mined information help or guide us in real life?

To answer our research questions, we have defined a methodology based on NLP techniques, such as a sentiment analysis and topic modelling. Our approach collects, ingests, processes, and analyses tweets for studying the sentiment and topics of interest. We have omitted the COVID-19 pandemic to be able to analyse the other viruses that afflict humans. We have observed that people mainly discuss malaria, influenza, and tuberculosis and are concerned about the disease itself. According to the sentiment analysis, fear, trust, and disgust are the three most dominant emotions. However, people have shown the trust emotion with respect to HIV/AIDS tweets.

The related works are discussed in Section 2, and we describe our study methodology in Section 3. Then, from Sections 4–9, we provide information on the data collection, data preprocessing, data exploration, vectorisation, sentiment analysis, and topic modelling. Finally, Section 10 concludes the paper.

2. Related Works

Numerous studies have already explored the analysis of pandemic events on social media by using machine learning techniques and natural language processing.

In the following, we are going to summarise studies that consider just one epidemic, such as COVID-19, Ebola, and influenza.

Zhang et al. [7] use five machine learning algorithms (decision tree, logistic regression, k-nearest neighbours, random forest, and support vector machine) based on the historically labelled coronavirus tweets dataset to build a sentiment classifier.

Sepúlveda et al. [8] present a real time tool, COVIDSensing, in which they use topic modelling and a sentiment analysis to analyse the socio-economic problems related to COVID-19 on Twitter, Really Simple Syndication, and Telegram.

Apart from the topic and sentiment, Imran et al. [9] also assign labels, such as geolocation, named entities, user types, and gender, for a dataset with two billion multilingual tweets about COVID-19. In order to geotag tweets, they use five meta-data attributes: tweet text, user location, user profile description, geo-coordinates, and place tags by geocoding and reverse geocoding. In the named-entity recognition task, they use named-entity recognition (NER) models to recognise eighteen different types of entities. Based on the user type, the first names of the identified personal accounts are employed for training a supervised machine learning classifier to classify gender.

When dealing with tweets with geolocation information, it is meaningful to estimate the users’ mobility. For example, if the length from the location of the first tweet identifier to the location of the second tweet identifier is larger than 100 m, Graff et al. [10] regard it as one trip. Then, they perform the same operation for all the users that published some messages on a specific day.

Cornelius et al. [11] present an interactive web platform to aggregate and visualise social media mining regarding COVID-19. They use the OntoGene Entity Recognizer for drug brand name detection. They detect URLs referring to preprint papers and estimate their popularity. In contrast, to address the general awareness of health issues, they use a Bidirectional Encoder Representation from Transformers (BERT)-based model to identify a personal health mention.

Andreadis et al. [12] explore the tweets spread about COVID-19 in Italy. They employ logistic regression and random forests to classify fake news or misinformation.

Other researchers [13,14] build models to identify the appearance of misinformation related to COVID-19 in different media.

Househ [15] focuses on the number of tweets and retweets related to Ebola and finds there is a correlation between electronic news media outlets and social media discussions.
Yousefinaghani et al. [16] analyse posts discussing avian influenza on Twitter. In detecting irrelevant tweets, they use an expectation-maximisation-based semi-supervised classifier to determine the class label of an unlabelled tweet.

Aramaki et al. [17] detect influenza on Twitter by a support vector machine to classify the negative and positive influenza tweets.

Santillana et al. [18] combine multiple influenza-like illnesses (ILI) activity estimates into a single prediction of an ILI by machine learning techniques, such as stacked linear regression, a support vector machine with radial basis function kernels, and AdaBoost with decision trees regression.

Twitter volumes can be regarded as a sign of real life. Gori et al. [19] create a relative increase indicator about the volume of tweets related to the vaccine of COVID-19 and investigated its tendency with real events.

Apart from tweets, retweet interactions can reflect a real-life community structure [20]. Although replies and quotes can also express a certain meaning, they are still ambiguous as compared to retweets [21].

Mahdikhani [22] combine the decision of a random forest, stochastic gradient descent, and logistic regression and generate a predictive model for the retweetability of posted tweets related to COVID-19. The result shows that tweets with a higher emotional intensity are more popular.

To improve the level of findability, Bellandi et al. [23] conduct an analysis on the COVID-19 scientific literature by combining different clustering methods (K-means, DBSCAN, agglomerative, MiniBatchkmeans, and BIRCH algorithms) with various vectorisation techniques (CountVectorizer, HashingVectorizer, TFIDF Vectorizer, word2vec, and doc2vec).

In the following, we are going to consider studies that use more than one epidemic, such as COVID-19 and influenza.

Alsudias and Rayson [24] monitor the COVID-19 pandemic and influenza epidemic by NLP techniques, including multilabel classification for finding infected people by a set of methods (such as multilabel k-nearest neighbours and BERT) and predicting the location for every infected person by a conditional random fields algorithm.

The above analyses are mainly based on a single pandemic. In our study, 11 pandemics are analysed and compared. In the sentiment part, previous researchers built classifiers according to the labelled data or assigned sentiments directly by an emotion lexicon. In this study, these two methods are combined and semi-supervised learning is used to label the sentiments. Furthermore, the topic modelling for a specific sentiment in a specific pandemic subset is built to investigate people’s attitude to these latent topics.

3. Methodology

Based on our research questions, we have defined a methodology (summarised in Figure 1) able to reflect on people’s opinions with respect to 11 epidemics using tweets.

Data have been collected from Twitter developer platform according to keywords related to 11 viruses: cholera, Ebola, HIV/AIDS, influenza, malaria, Spanish influenza, swine flu, tuberculosis, typhus, yellow fever, and Zika. We have employed the tweepy package [25] to scrawl data every Thursday at 9 a.m. Just English tweets have been considered, and retweets and replies have been filtered out. The considered period goes from 3 March to 15 September 2022, allowing to collect 369,472 tweets.

In the original dataset, there are the following columns: datetime is the time of tweet posting; tweet id and author id are the unique identifiers of a tweet and its author, respectively; original text is the content of the tweet; retweet count, replay count, and like count show the interaction with a specific tweet; geo is the geolocation information of the tweet (if it is reported in the tweet).
After obtaining the dataset for each epidemic and merging them together (de-duplication according to tweet id), it is essential to preprocess the text. The original texts have been cleaned by removing various noises, such as emojis and special characters, through the usage of natural language processing tasks. The cleaning steps are described in the following list:

- Removing user names, hashtags, URLs, non-ASCII characters, numbers, punctuation, and special characters by using regular expressions;
- Making lowercase;
- Removing stop words that are common words (such as the, is, at, which, and on) but have no real meaning by using the spacy package [26];
- Performing lemmatisation to recover a word to its original form (e.g., transforming ate into eat) by using WordNetLemmatizer package [27];
- Correcting spelling errors by using the autocorrect package [28].

Once data are preprocessed, we have started to explore the findings by using basic visualisations. We have, e.g., plotted the dataset distribution, the number of tweets over time, the word frequencies, and the tweets’ geolocation.

We have transformed text into numerical representation, because the computer is not able to understand text directly. NLP provides several vectorisation techniques: some are based on sentences or they produce sentence vector representation directly, e.g., bag of words (BOW) [29] and term frequency (TF)–inverse document frequency (IDF) [30]. A bag of words is a simple representation of text that describes the occurrence of words within a document. TF-IDF evaluates how relevant a word is to a document in a collection of documents.

Figure 1. Methodology overview.
Some others NLP techniques focus on words to build word vectors, e.g., Word2Vec [31] and FastText [32]. Word2Vec is an unsupervised learning technique that uses a shallow, two-layer neural network to train and reconstruct linguistic contexts of words: this technique can utilise continuous bag of words (CBOW) or continuous skip-gram, where the model uses the current word to predict the surrounding window of context words. FastText is a word embedding and text classification method sourced by Facebook in 2016 that often achieves comparable accuracy to deep networks.

According to these vectorisations’ characteristics, different machine learning and natural language processing techniques can be applied. Clustering based on word embedding identifies similarity or dissimilarity between observations according to their distances. We have considered Word2Vec and FastText for word embedding and $k$-means to calculate the Euclidean distance between observations [33].

Furthermore, we have used Latent Dirichlet Allocation (LDA)-based topic modelling [34] with bag of words to find the latent topics. Sometimes one tweet discusses more than one topic. In natural language processing, the topic modelling approach represents an unsupervised learning method to find topic distribution in corpus. This solution can be performed by using the Latent Dirichlet Allocation (LDA) model, i.e., a Bayesian probabilistic model, that is used to determine the latent topic and its probability distribution for each document in the corpus. LDA leverages the bag-of-words (BOW) model.

We have also performed sentiment analysis on TF-IDF [35] by using uni-gram and bi-gram to consider the order of words and to explore the sentiment distribution.

### 4. Exploring Data

In this section, we speculate about the data information. Figure 2 shows the composition of our datasets. Malaria accounts for about 28.3% of the whole data, while influenza with 17.6% and tuberculosis with 13.8% are in the second and third positions, respectively. We can observe that the discussion about typhus is the smallest one, just 0.7%.

![Figure 2. The composition of the 11 datasets.](image)

Figure 3 shows the word cloud for the total dataset. The keywords about viruses have been removed to exclude their influence on the resulting plot. We can observe that twitter users are concerned about people’s health issues in relation to viruses around the world. Terms, i.e., people, health, vaccine, and disease, have a higher frequency with respect to others, such as work.
For different viruses, word clouds and bar plots with the first 20 words have been created. In this paper, we have included the main plots. Figure 4 is for Ebola, which highlights the terms Congo and outbreak. On the one hand, Ebola is the name of a river in the northern part of the Democratic Republic of Congo, in which an unknown virus came and killed people in 55 villages along the Ebola River in 1976. On the other hand, on 23 April 2022, the World Health Organization [36] issued a statement that Mbandaka city, a north-western Equateur province capital, in the Republic of Congo, found a person suffering from Ebola haemorrhagic fever, and the country’s health department declared an outbreak for a new round of Ebola. Meanwhile, the frequency of Marburg is very high. The Marburg virus is characterised by the same symptoms and transmission routes as the Ebola virus disease. The Marburg virus and Ebola virus belong to the Filoviridae family. Marburg has a high frequency, because on 1 July, Ghana had confirmed the first two fatal cases of the Marburg virus disease.

Figure 5 is for cholera which is often associated with water pollution. In this case, we have observed that the water term appears frequently. Terms, such as Mariupol and Ukraine, also have a high frequency. Mariupol is a port city in Ukraine, which has been almost destroyed by the current war. Water has mixed with sewage, and according to the BBC news [37], there could be the risk of a major cholera outbreak.

Young people are often at high risk of HIV/AIDS. There are several tweets discussing prevention behaviour or awareness for youth. On 10 March 2022, there is the National Women and Girls HIV/AIDS Awareness Day, and it is possible to observe various tweets about this argument.

Influenza tends to be associated with avians and pigs. We have observed various tweets that include bird, flock, and so on.
Malaria is mainly spread by mosquitoes. The mosquito term occupies a high percentage. The child word also has a high frequency, which means that a malaria infection for children is of great concern.

Spanish flu is a disaster in history, known as the 1918 influenza pandemic. Terms such as year, time, and history have a high frequency.

For tuberculosis, children and elderly people are mainly interested. The meningitis term also has a high proportion, because tuberculous meningitis is one of the typical complications of tuberculosis. This disease mostly occurs in children under 5 years of age, and the elderly are also a susceptible population.

The Queensland term has a high proportion in the typhus dataset. There are many types of typhus virus and the Queensland tick typhus is one of them [38]. It is a zoonotic disease caused by the bacterium Rickettsia australis.

In the swine flu dataset, terms such as war, Ukraine, and Russia appear and account for a significant percentage. This is related to the latest international events and conflicts. Furthermore, in the same dataset, terms such as people and vaccine appear and account for a significant percentage. It is also closely related to terms such as bird and poultry.

Kenya, outbreak, and die appear frequently related to yellow fever. According to the news, on 5 March 2022, the Kenyan Ministry of Health declared an outbreak of yellow fever in the country [39].

Like malaria, Zika tends to be associated with the term mosquito. It is also associated with dengue, which is another infectious disease caused by mosquitoes.

The #malaria hashtag has the highest number as well as the highest number of malaria tweets collected. Although this study does not collect tweets for the COVID-19 keyword, there are a lot of hashtags about it. It means that COVID-19 always accounts for a significant proportion of the discussion about pandemics. Apart from hashtags related to specific viruses, there are also other kinds of hashtags: #EndTigraySiege, #Mekelle, and #Ethiopia. The Tigray region is the northernmost regional state in Ethiopia. Mekelle is the capital of the Tigray region. Because of the influence of the civil war in Ethiopia [40], Tigray has been under siege for a long period. There is not enough food and medicine supply, which leaves hundreds dying daily and millions risking death. So, there are many tweets appealing that Tigray needs urgent assistance to save lives. As for #AyderReferralHospital, due to limited medical resources, there are several patients infected due to pandemics who do not receive treatment. For example, at the Ayder Referral Hospital [41], babies with meningitis and tuberculosis and a 14-year-old boy with HIV have been turned away.

Figure 6 shows the number of tweets over time for different virus subsets. There are approximately 1875 relevant tweets posted each day during this period. It is interesting to compare the number of tweets with the facts in our daily life. Figure 6 shows several peaks that can be put in relation to specific events in real life, e.g., 24 March was World Tuberculosis Day and 25 April was World Malaria Day. The peak for HIV/AIDS is a little higher, i.e., 1034, on 10 March 2022, because there was National Women and Girls HIV/AIDS Awareness Day. The peak of the tuberculosis tweets was 1238 on 29 April.
2022. Furthermore, the 24–30 April 2022, week was World Immunization Week 2022, when people’s discussion about viruses increased for tuberculosis.

![Figure 6](image1.png)

Figure 6. The number of tweets over time for the various epidemics.

For other peaks, there are corresponding events: on 23 June 2022, commonwealth leaders recommitted to ending malaria; on 21 July, Gavi, the Vaccine Alliance, funded 160 million EUR to increase malaria vaccine access in Africa; on 30 August 2022, a total of 75 cases of Cholera were reported in Nepal from Kathmandu, Lalitpur, Bhaktapur, Nuwakot, and Dhading cities; and on 8 September, new phase 2b findings showed the Oxford malaria vaccine maintains a high level of protection.

Figure 7 shows an earth map with the location of some tweets by considering the attribute place id, which is a unique identifier (ID) for the location on Twitter. With this ID, we have been able to obtain detailed information about the place type, the full name of this place, and the country to which it belongs.

![Figure 7](image2.png)

Figure 7. The location of tweets related to Malaria.

Investigating the distribution of tweets around the world is an interesting point. In this case, let us just focus on geolocated tweets by filtering out those non-geolocated observations. The total size of the original dataset is 369,472, and after filtering, our research obtains just 6863 tweets located.

For malaria, most of the geolocated tweets are from Nigeria, India, Uganda, the United States (US), and Kenya. For tuberculosis, tweets from India account for the biggest proportion. For influenza and HIV/AIDS, tweets from the United States are the most frequent. As for other viruses, there is no significant pattern in distribution. Because of the limit of the size, the result is not accurate enough. In the research, just English tweets are considered. So, it is not strange that the United States always has a high proportion of data.
Figure 8 shows a distribution of the breath symptom among the 11 viruses. To identify which symptoms account for the highest percentage in a given dataset, we have calculated their frequency and distribution in different viruses and represented them by using bar plots (sorted by descending order). The considered pandemics present the following symptoms: breath, fever, diarrhea, headache, rash, cough, chill, fatigue, coma, death, jaundice, muscle pain, and a weakness illness.

Different viruses have a different size of observations (e.g., the malaria dataset is the biggest one, but the typhus one has the smallest size), which means that for the same symptom, the virus dataset with the biggest size tends to have the largest frequency of a given symptom. To avoid that the size of each dataset influences the symptom distribution, every frequency has been divided by the size of the vocabulary of its corresponding data source.

Figure 8. Bar plot of breath problem.

For the breath problem (see Figure 8), tweets related to typhus appear more often than the other viruses. Similarly, compared to the other viruses, a discussion about fever accounts for a higher proportion in the yellow fever dataset. Compared with the others, a rash and diarrhea are typical symptoms for typhus and cholera, respectively; a cough is also widely discussed in the tuberculosis dataset; death has a higher proportion in the Spanish flu, and according to history, there are exactly so many dead cases because of the Spanish flu; jaundice is the feature of yellow fever; and for other symptoms, there is no clear distinction among the virus datasets.

These considerations are drawn from the point of statistics instead of the perspective of medicine. Of course, we can explore the different symptoms discussion frequency within the same virus dataset.

Figure 9 shows the retweets, replies, and likes frequencies for the 11 pandemics. The number of retweets, replies, and likes can be regarded as a symbol of interaction on social media. In general, the number of likes is much more than the retweets and replies. Specifically, malaria has the most likes, replies, and retweets. It means that tweets about malaria are popular on Twitter.
5. Vectorisation

In this study, four vectorisations have been considered: the bag of words (BOW), the TF-IDF vector based on uni-grams and bi-grams, Word2Vec, and FastText.

In word embedding, it is essential to specify dimensionality. According to some articles, there are some methods to find the optimal size of a word vector in which the most important step is to evaluate the performance of the word embedding. For example, Yin and Shen [42] introduce a Pairwise Inner Product loss function. In our study, we have considered the method of Faruqui and Dyer [43] which evaluates the performance of the word embedding by calculating the Spearman correlation between the similarity score (regarded as the ground truth) and the cosine similarity in a vector space for matched pairs of words. The dimensionality of the word vector changes from 100 to 300. We have computed the optimal dimensionality which maximises the correlation. Figure 10 shows that 170 and 110 are the best numbers of dimension for Word2Vec and FastText, respectively.

![Figure 9. Retweets, replies, and likes.](image)

![Figure 10. Dimension for Word2Vec on the left and FastText on the right.](image)

After ensuring the dimensionality and obtaining the word embeddings, the interpretation of the dimensions is always a difficult task to deal with. Tsvetkov et al. [44] exploit an existing semantic resource—SemCor—to interpret individual vector dimensions. SemCor is an English corpus with 41 kinds of supersense annotations, such as NN.ANIMAL and VB.MOTION. Based on SemCor, they construct 4199 linguistic word vectors with 41 interpretable columns, which are called linguistic property vectors. Then, they take an alignment between the word vector dimensions and the linguistic dimensions which
maximises the cumulative Pearson’s correlation between the aligned dimensions of the two matrices.

Finally, for every document, according to Word2Vec and FastText, new sentence vectors are built by calculating the mean vector of all the words within a sentence. Of course, there is the drawback that we may obtain the same interpretation for different columns because of the size of the linguistic property.

6. Clustering

In this study, we have used the \( k \)-means method for clustering tweets. We have selected the number of clusters by considering a range of values between 2 and 10. We have also calculated the sum of the distances of the data coordinates (i.e., the silhouette score [45]) from the cluster centroids for every \( k \)-means model: this value decreases when the number of clusters increases. Figure 11 shows that 8 is the best \( k \) choice for Word2Vec and FastText.

![Figure 11. The choice of number of clusters—Word2Vec on the left side and FastText on the right side.](image)

In order to interpret the results, the top 30 words are listed according to their frequency in the different clusters. Tables 1 and 2 summarise the clusters based on Word2Vec and FastText, respectively. To understand the similarity between the clusters, we have considered the adjusted Rand index (ARI) [46]. Its domain is \([-1, 1]\), and the closer the value is to 1, the more similar they are. We have obtained an ARI value equal to 0.426, which means that the two clusters are relatively similar.

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>Interpretation</th>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster0</td>
<td>medical treatment</td>
<td>vaccine, disease, case</td>
</tr>
<tr>
<td>cluster1</td>
<td>people’s health situation</td>
<td>old, baby, people</td>
</tr>
<tr>
<td>cluster2</td>
<td>cause of virus</td>
<td>mosquito, world</td>
</tr>
<tr>
<td>cluster3</td>
<td>war and conflict</td>
<td>Ukraine, war, refugee</td>
</tr>
<tr>
<td>cluster4</td>
<td>deaths</td>
<td>people, die, kill</td>
</tr>
<tr>
<td>cluster5</td>
<td>avian</td>
<td>bird, county, flock</td>
</tr>
<tr>
<td>cluster6</td>
<td>pandemic outbreak in Congo</td>
<td>Congo, outbreak, crisis</td>
</tr>
<tr>
<td>cluster7</td>
<td>people’s activity such as National Women and Girls HIV/AIDS Awareness Day</td>
<td>awareness, world</td>
</tr>
</tbody>
</table>
Table 2. Interpretation of clustering on FastText.

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>Interpretation</th>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster0</td>
<td>people’s activity such as National Women and Girls HIV/AIDS Awareness Day</td>
<td>people, awareness, national</td>
</tr>
<tr>
<td>cluster1</td>
<td>medical treatment</td>
<td>vaccine, people, mask</td>
</tr>
<tr>
<td>cluster2</td>
<td>fight against viruses</td>
<td>world, health, fight</td>
</tr>
<tr>
<td>cluster3</td>
<td>pandemic situation in a certain area</td>
<td>outbreak, Congo, city</td>
</tr>
<tr>
<td>cluster4</td>
<td>avian</td>
<td>bird, county, flock</td>
</tr>
<tr>
<td>cluster5</td>
<td>deaths</td>
<td>die, people, time</td>
</tr>
<tr>
<td>cluster6</td>
<td>war and conflict</td>
<td>Ukraine, Russia, kill</td>
</tr>
<tr>
<td>cluster7</td>
<td>cause of viruses</td>
<td>mosquito, world</td>
</tr>
</tbody>
</table>

The clustering method has been applied to the overall dataset. However, to obtain a more concrete result, it is also possible to take the clustering on a specific pandemic dataset.

7. Topic Modelling

In our study we have applied the LDA model that performs the following operations several times to create documents: First, it selects one of the predefined topics with a certain probability, and then selects a word under that topic with a certain probability. Assume there are $M$ documents with $K$ topics. Each document (length $N$) has its own topic distribution that is polynomial with the parameters of the polynomial distribution that obey the Dirichlet distribution and are $\alpha$. Each topic has its own word distribution that is a multinomial distribution with the parameters of the multinomial distribution that obey the Dirichlet distribution and are $\beta$. For the creation of the $n$-th word in a given document, a topic is first selected from the topic distribution of that document, and then a word is selected from the word distribution corresponding to that topic. This generation process is repeated until all $M$ documents complete the above process.

We have used the gensim library [47] to perform the LDA. Apart from the input of the sentence vector (bag of words) and the dictionary (id and word), it is essential to specify the number of topics. In order to find the optimal values, topic coherence is employed as the indicator to measure the performance of the model. It is meaningful to calculate the frequency of the co-occurrence of the words belonging to the same topic in the corpus. Topic coherence does just that. The gensim library offers several different measures of topic coherence, and the main difference is the definition of “co-occurrence”, where $c_v$, $c_{uci}$, $u_{mass}$, and $c_{npmi}$ are optional methods. Here, the number of topics varies from 1 to 20 to find an optimal value that maximises the $c_v$ coherence.

Figure 12 shows that seven is the optimal value of the number of topics. The model with seven topics can obtain a relatively high coherence score.

Figure 12. Determining optimal number of topics.
The LDA model can be used to extract the latent topics. Its results can be visualised with the pyLDAvis library [48], which is an open-source package in Python, to interactively present the results of the LDA. Figure 13 shows the top 30 words and the main topics that can be explained as follows: Topic1—malaria; Topic2—cholera in Mariupol, Ukraine; Topic3—tuberculosis; Topic4—stop HIV/AIDS; Topic5—influence or flu; Topic6—new cases infected; and Topic7—other diseases, such as cancer. This result is too general because there are 11 different epidemics sources.

Figure 13. Topic modelling results based on LDA obtained with the pyLDAvis library.

8. Sentiment Analysis

In this research, we have combined a lexicon and semi-supervised learning techniques to perform a sentiment analysis.

There are no emotion or sentiment labels in our data; therefore, we have used a lexicon to label emotion. Furthermore, we have considered an emotion classifier, based on the National Research Council Canada (NRC) Affect Intensity Lexicon [49], available in Python with the emotion-nrc-affect-lex package [50], that identifies emotions and computes an aggregated score for each emotion. This classifier uses a lexicon that has around 10,000 entries for eight emotions: fear, anger, anticipation, trust, surprise, sadness, disgust, and joy. Specific rules have been defined to label the sentiments: for every tweet and corresponding emotion distribution, we have selected the emotion with the highest weighted emotion score; if there are no emotions, because no word matched, we have assigned the neutral sentence label. According to this approach, each tweet can have a sentiment assigned.

Once the sentiments are labelled, we performed semi-supervised learning. The dataset, as shown in Figure 14, has been divided into two parts: the training dataset (80%) and the test dataset (20%). The training data have been used to build the classifier and the testing data are used to measure the performance. Particularly, for the training data, 80% of the labels of the data are removed and the remaining labels are regarded as the ground truth. We have defined the classifiers by considering the labelled training data and using logistic regression (LR), a multinomial Bayes (MNB) model, and random forest (RF) based on the TF-IDF vector. Then, we have measured their performances on our test dataset, which is summarised in Table 3. According to the accuracy value of 0.80, we have selected the logistic regression model.
We have also applied a self-training approach that belongs to the semi-supervised machine learning algorithms, as it uses a combination of labelled and unlabelled data to train the model. The idea behind the self-training approach consists of:

- Using the labelled data to train the first supervised model, such as the logistic regression one;
- Using the model to predict the class of the unlabelled data;
- Selecting the tweets that satisfy the predefined criteria (e.g., with a prediction probability of 96% or belonging to the top 10 observations with the highest prediction probability);
- Combining these pseudo-labels with the labelled data;
- Using the labels and pseudo-labels to train a new supervised model;
- Making predictions again and adding the new observations to the pseudo-labelled pool;
- Iterating these steps until no other unlabelled observations satisfy the pseudo-labelling criterion, or when the specified maximum number of iterations is reached;
- Finally, defining an adjusted or improved logistic regression classifier (whose accuracy is 80%) that labels the sentiment of all the observations.

Figure 15 shows that fear, trust, and disgust are the three most dominant emotions. In detail, 38.4% of the tweets have the fear emotion to pandemics and 16.5% of the tweets contain the disgust mood. At the same time, there are 20.5% of tweets which express a trust emotion.
Figure 15 shows the sentiment distribution in different pandemic datasets: in cholera, Spanish flu, and yellow fever, fear dominates among the emotions; in Ebola, influenza, tuberculosis, typhus, and Zika, the trust emotion occupies a significant proportion, although fear is the biggest one; in HIV/AIDS, the trust emotion accounts for a larger proportion than fear; and in malaria and swine flu, disgust has the highest frequency, followed by fear and trust. According to our findings, users on social media have negative emotions to pandemics, but in some cases, e.g., Ebola, influenza, tuberculosis, typhus, and Zika, people still keep a certain positive attitude, such as for HIV/AIDS, where people show a trust emotion.

9. Combining Topic Modelling and Sentiment Analysis

The previous topic modelling and sentiment analysis are too general. In order to explore users’ attitudes towards specific topics, the LDA model has been built to find the latent topics with respect to pandemics and emotions.

Tables 4 and 5 show that the fear emotion is dominant for 8 out of 11 viruses, such as Ebola, cholera, influenza, Spanish flu, swine flu, tuberculosis, typhus, yellow fever, and
Zika. Furthermore, the trust emotion is prevalent in HIV/AIDS and tuberculosis, while the disgust emotion is common in cholera and malaria. In five cases, the main topic is related to COVID-19. Two topics report special events, such as the World Tuberculosis Day and the National Women and Girls HIV/AIDS Awareness Day. In many cases, the topics include one possible reason for the virus, such as the bite of a mosquito for malaria and water pollution for cholera.

Table 4. Topics for pandemic and emotion pair: Ebola, cholera, influenza, and Spanish flu are listed with the fear emotion; HIV/AIDS is with the trust emotion; malaria and swine flu show topics with the disgust emotion.

<table>
<thead>
<tr>
<th>Pandemic</th>
<th>Dominant Emotion</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebola</td>
<td>fear</td>
<td>Topic1—COVID-19&lt;br&gt;Topic2—Information about the virus&lt;br&gt;Topic3—Ebola outbreak in Congo&lt;br&gt;Topic4—Political issues in Congo such as scandal</td>
</tr>
<tr>
<td>HIV/AIDS</td>
<td>trust</td>
<td>Topic1—Research findings&lt;br&gt;Topic2—National Women and Girls HIV/AIDS Awareness Day&lt;br&gt;Topic3—High-risk or susceptible groups&lt;br&gt;Topic4—Help and encourage from experts</td>
</tr>
<tr>
<td>Malaria</td>
<td>disgust</td>
<td>Topic1—Treatment such as vaccine&lt;br&gt;Topic2—People’s reaction&lt;br&gt;Topic3—The spreading of Malaria: the bite of mosquito&lt;br&gt;Topic4—Research about Malaria&lt;br&gt;Topic5—The cause of Malaria: the falciparum parasite&lt;br&gt;Topic6—Organisations or figures&lt;br&gt;Topic7—Usage of drug&lt;br&gt;Topic8—Medical treatment’s achievement</td>
</tr>
<tr>
<td>Cholera</td>
<td>fear</td>
<td>Topic1—COVID-19&lt;br&gt;Topic2—The risk of Cholera outbreak in Mariupol, Ukraine&lt;br&gt;Topic3—The cause of Cholera: water pollution</td>
</tr>
<tr>
<td>Influenza</td>
<td>fear</td>
<td>Topic1—COVID-19&lt;br&gt;Topic2—Avian&lt;br&gt;Topic3—The risk of the virus&lt;br&gt;Topic4—New cases and deaths</td>
</tr>
<tr>
<td>Spanish flu</td>
<td>fear</td>
<td>Topic1—COVID-19&lt;br&gt;Topic2—Avian&lt;br&gt;Topic3—Million deaths in the history&lt;br&gt;Topic4—Protective measures such as wearing masks</td>
</tr>
<tr>
<td>Swine flu</td>
<td>disgust</td>
<td>Topic1—COVID-19&lt;br&gt;Topic2—The connection with avian influenza&lt;br&gt;Topic3—Influence on the world&lt;br&gt;Topic4—Animals such as swine, monkey, and so on</td>
</tr>
</tbody>
</table>
Table 5. Topics for pandemic and emotion pair: tuberculosis, typhus, yellow fever, and Zika are listed with the fear emotion; tuberculosis also shows topics for the trust emotion.

<table>
<thead>
<tr>
<th>Pandemic</th>
<th>Dominant Emotion</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuberculosis</td>
<td>fear</td>
<td>Topic1—COVID-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic2—Viral resistance to drugs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic3—Other diseases, such as cancer and diabetes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic4—Call to fight the virus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic5—Achievement of treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic6—Infection in the prison</td>
</tr>
<tr>
<td></td>
<td>trust</td>
<td>Topic1—World Tuberculosis Day to raise people’s awareness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic2—New information from research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic3—Medical system</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic4—Collaboration and campaign around the world</td>
</tr>
<tr>
<td>Typhus</td>
<td>fear</td>
<td>Topic1—Deaths</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic2—The outbreak of disease and war</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic3—Vaccine</td>
</tr>
<tr>
<td>Yellow fever</td>
<td>fear</td>
<td>Topic1—Vaccine</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic2—Outbreak in Kenya</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic3—Deaths</td>
</tr>
<tr>
<td>Zika</td>
<td>fear</td>
<td>Topic1—Infection caused by mosquito</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic2—Dengue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic3—The outbreak of Zika</td>
</tr>
</tbody>
</table>

10. Discussion and Conclusions

In this work, we have used different natural language processing and machine learning techniques to explore the pandemics’ information on Twitter social media. We have excluded COVID-19 in order to avoid having unbalanced data for the other viruses. The findings support us to answer our research questions.

RQ1. Which pandemics have more discussion on social media? What is the trend of these discussions over time?

Despite COVID-19 being excluded from our analysis, the collected data show that COVID-19 is accounted for by people’s discussions according to the frequency of hashtags and topic modelling. Furthermore, discussions about malaria, influenza, and tuberculosis occur the most on Twitter, while the number of tweets related to typhus is the smallest one. We have also observed that malaria, influenza, and tuberculosis are the most popular according to the number of retweets, replies, and likes. From our understanding, the main reasons are: (1) the presence of two special days about malaria and tuberculosis; (2) influenza is a very general term and has many variations, such as Spanish flu and swine flu; and (3) nowadays, typhus is actually considered a rare disease [51].

RQ2. What are people’s concerns related to these pandemics?

Our study deals with this question from several different points of view. Firstly, by calculating the frequency of words, we have identified the top 30 words or hashtags to explore people’s concerns. Secondly, after the vectorisation of each tweet message, we have computed the distance between the vectors to explore the observations’ similarity by using $k$-means and then interpreted every cluster. Finally, according to the topic modelling, we have determined the latent topic for every tweet. We have understood that some people’s concerns are related to the disease itself while others are related to politics and war.

RQ3. What are their attitudes or emotions to these epidemics?
From the result of the sentiment analysis, fear, trust, and disgust are the three most dominant emotions. Specifically, we have presented the emotion distribution for every pandemic.

RQ4. How can the mined information help or guide us in real life?

According to the last section, we have found that people are scared of most of the pandemics. The frightening topics are often related to the cause of the pandemics and their influence on human beings and society. It is worth highlighting that people have a fear emotion to several medical treatments, such as wearing masks or taking a vaccine, although these measures are effective in controlling pandemics. In order to eliminate people’s fear, it is important for the governments or departments concerned to focus on propaganda to make people understand the benefit of these treatments. People are equally afraid of some human activities, such as wars, biological experiments, and some political issues, which have a strong correlation with pandemics. In order to reduce the impact of these pandemics as soon as possible, we should call for peace and protect our environment, because apart from the influence of nature, there are some human factors in the outbreaks of these pandemics. Our findings show that there are also positive emotions or attitudes. For example, people have a trust emotion in tweets related to HIV/AIDS.

Author Contributions: Conceptualisation, E.R. and Z.Q.; methodology, E.R. and Z.Q.; software, Z.Q.; validation, E.R. and Z.Q.; formal analysis, E.R. and Z.Q.; investigation, Z.Q.; data curation, Z.Q.; writing—original draft preparation, E.R.; writing—review and editing, E.R.; visualisation, Z.Q.; supervision, E.R.; project administration, Z.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Description of the Considered Pandemics

Cholera is a bacterial infection by some strains of the bacterium vibrio cholerae, which results from eating or drinking contaminated food and water. The typical symptom is a large amount of watery diarrhea that lasts a few days.

Ebola is a viral haemorrhagic fever caused by Ebolaviruses. The specific symptoms are sore throat, fever, headaches, and muscle pain. These are usually followed by vomiting, rash, diarrhea, and decreased liver and kidney function.

HIV/AIDS is a human immunodeficiency virus infection and acquired immunodeficiency syndrome. This virus attacks the immune system. Swollen lymph nodes, fever, and headaches are typical symptoms.

Influenza, also known as the flu, is a disease caused by the influenza virus that infects the respiratory tract. The most common symptoms are fever, sore throat, runny nose, headache, cough, and general malaise.

Malaria is a parasitic disease transmitted by mosquitoes. Typical symptoms caused by malaria are fever, fatigue, chills, headache, and vomiting; in severe cases, it can cause jaundice, seizures, coma, or even death.

Spanish influenza, or the 1918 flu pandemic, was an unusual deadly influenza pandemic that broke out between January 1918 and April 1920. The main symptoms were sore throat, headache, fever, and mucosal haemorrhage, but it led to death if it became severe.

Swine flu is an infection caused by several types of swine influenza viruses. The swine influenza virus is common throughout pig populations. If its transmission causes a human flu, it is called a zoonotic swine flu. The symptoms of a zoonotic swine flu are similar to
influenza. In general, the symptoms are chills, breath shortness, sore throat, muscle pains, headache, fever, coughing, weakness, and general discomfort.

Tuberculosis is an infection caused by the microbacterium tuberculosis bacteria that mainly affects the lungs. Its symptoms are a persistent cough (lasting more than 14 days) and fever.

Typhus is an infectious disease caused by bacteria that is transmitted to humans by the bite of fleas and ticks. Common symptoms include fever, headache, and rash.

Yellow fever is a viral infection characterised by severe fever and jaundice. It is caused by the yellow fever virus and is spread by the bite of an infected mosquito.

Zika is a viral infection transmitted by the Aedes aegypti mosquito. Common symptoms include fever, rash, and conjunctivitis.

References
9. Imran, M.; Qazi, U.; Ofili, E. TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels. *Data* 2022, 7, 8. [CrossRef]

22. Mahdikhani, M. Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic. *Int. J. Inf. Manag. Data Insights* 2022, 2, 100053. [CrossRef]


47. gensim. gensim 4.2.0. Available online: https://pypi.org/project/gensim/ (accessed on 16 October 2022).


50. NRC. Emotion-Nrc-Affect-Lex 0.0.3. Available online: https://pypi.org/project/emotion-nrc-affect-lex/ (accessed on 16 October 2022).