 doi:10.20944/preprints202211.0005.v1

Article

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, Bologna, 40126, Italy; qinzhikang168@163.com (Z.Q.)
[2] INFN CNAF, Bologna, 40126, Italy
* Correspondence: elisabetta.ronchieri@cnaf.infn.it (E.R); 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 since 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 analyzed 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 the 3rd of March to the 15th of September, 2022. Our analysis has started with the collection of comments composed of unstructured texts on which we have applied natural language processing solutions. Afterward, we have employed topic modelling and sentiment analysis techniques to obtain a collection of people's concerns and attitudes toward 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; HIV/AIDS; Influenza; Malaria; Spanish influenza; Swine flu; Tuberculosis; Typhus; Yellow fever; and 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, this COVID-19 is just one of the pandemics that cause suffering on our planet. Social media represent a valid instrument to understand public perceptions during a pandemic, providing some guidelines to governments and medical organizations.

In this work, we have collected tweets - written from the 3rd of March to the 15th September - 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 service in the world. As proved in previous literature, social media have been becoming 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 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 analyze epidemic-related messages on Twitter. By using several machine learning techniques, we aim at answering the following research questions:

 doi:10.20944/preprints202211.0005.v1

RQ1 Which pandemics have more discussion in 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 What 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 sentiment analysis and topic modelling. Our approach collects, ingests, processes, and analyses tweets for studying 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 about 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 trust emotion with respect to HIV/AIDS tweets.

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

2. Related Works

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

In the following, we are going to summarize 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 neighbors, random forest, and support vector machine) based on the historically labeled coronavirus tweets dataset to build a sentiment classifier.

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

Apart from 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 recognize eighteen different types of entities. Based on 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 users' mobility. For example, if the length from the location of first tweet identifier to the location of second tweet identifier is larger than 100 meters. Graff *et al.* [10] regard it as one trip. Then they perform the same operation for all the users that published some message on a specific day.

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

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

Other researches [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.

 doi:10.20944/preprints202211.0005.v1

Yousefinaghani *et al.* [16] analyze posts discussing avian influenza on Twitter. In detecting irrelevant tweets, they use expectation-maximization based semi-supervised classifier to determine the class label of unlabeled tweet.

Aramaki *et al.* [17] detect influenza on Twitter by Support Vector Machine to classify negative and positive influenza tweets.

Santillana *et al.* [18] combine multiple influenza-like illnesses (ILI) activity estimates into a single prediction of ILI by machine learning techniques, such as stacked linear regression, 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 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 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 higher emotional intensity are more popular.

To improve the level of findability, Bellandi *et al.* [23] take analysis on COVID-19 scientific literature by combining different clustering methods (K-means, DBSCAN, agglomerative, MiniBatchkmeans and BIRCH algorithms) with various vectorizations 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 neighbors, and BERT) and predicting location for every infection person by conditional random fields algorithm.

Above analyses are mainly based on a single pandemic. In our study, 11 pandemics are analyzed and compared. In the sentiment part, previous researches built classifiers according to labeled data or assign sentiments directly by emotion lexicon. In this study, these two methods are combined and a semi-supervised learning is used to label sentiments. Furthermore, topic modeling for specific sentiment in 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 (summarized 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 am. Just English tweets have been considered and retweets, and replies have been filtered out. The considered period goes from March 3rd, 2022 to September 15th, 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 getting 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 us-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

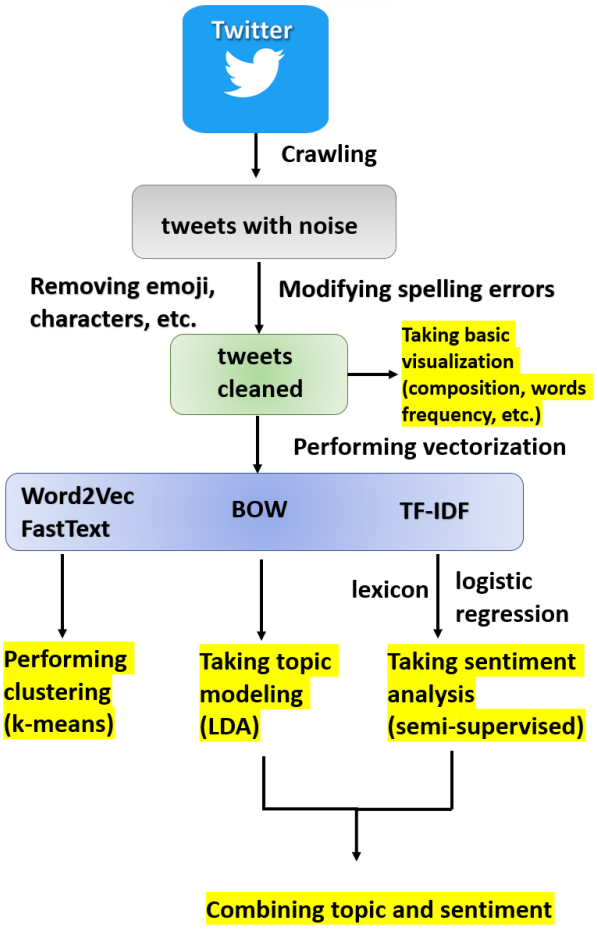


Figure 1. Methodology overview

age of natural language processing tasks. The cleaning steps are described in the following list:

- removing user names, hashtags, URLs, non-ASCII characters, numbers, punctuations, 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 lemmatization 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 preprocessed data we have started to explore the findings by using basic visualizations. 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 vectorization 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.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

can utilize 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 vectorizations' 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 modeling [34] with bag of words to find the latent topics. Sometimes one tweet discusses more than one topics. In natural language processing, the topic modeling 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 data information. Figure 2 shows the composition of our datasets. Malaria accounts 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%.

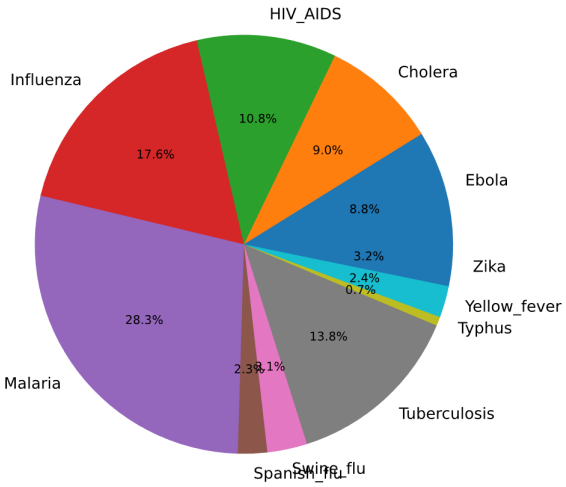


Figure 2. The composition of the 11 datasets

Figure 3 shows the word cloud for the total dataset. Keywords about viruses have been removed to exclude their influence in 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 like 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 that highlight 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 April 23rd, 2022, the World Health Organization [36] issued a statement that Mbandaka city, a northwestern Equateur province capital, in the Republic of Congo found a person suffering from ebola

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

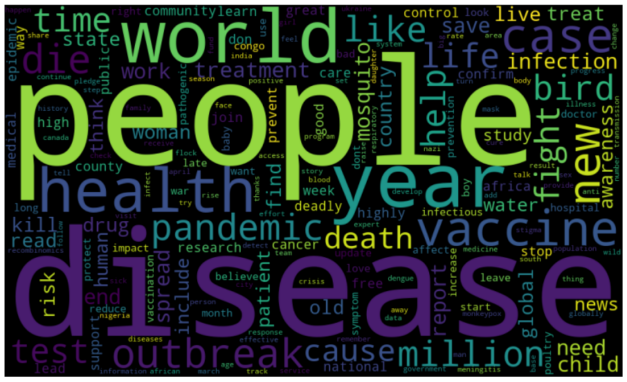


Figure 3. Word cloud once removed keywords about viruses

hemorrhagic 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 characterized 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 July 1st, Ghana had confirmed the first two fatal cases of the marburg virus disease.

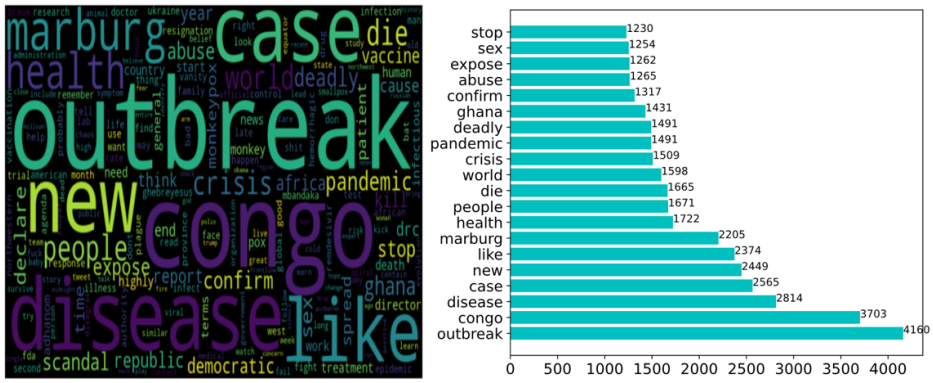


Figure 4. Word cloud and bar plot for ebola

Figure 5) is for cholera that is often associated with water pollution. In this case, we have observed that the water term appears frequently. Terms, like 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.

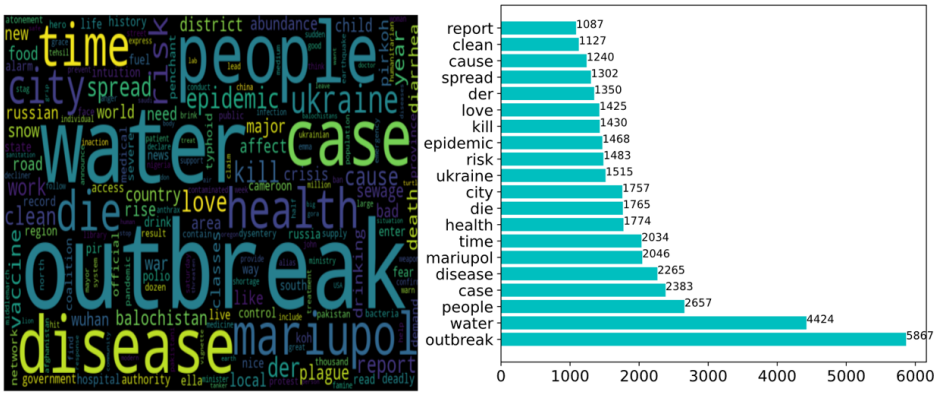


Figure 5. Word cloud and bar plot for cholera

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

Young people are often at high risk of HIV/AIDS. There are several tweets discussing prevention behavior or awareness for youth. On March 10th 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 malaria infection for children is of great concern.

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

For Tuberculosis, children and elderly people are mainly interested in. 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 like 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 like people and vaccine appear and account for a significant percentage. It is also closely related to terms like bird and poultry.

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

Like malaria, zika tends to be associated with term mosquito. It is also associated with dengue, that is another infectious disease caused by mosquito.

Figure 6 shows the frequency of the top 30 hashtags. The #malaria hashtag has the highest number as well as the 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.

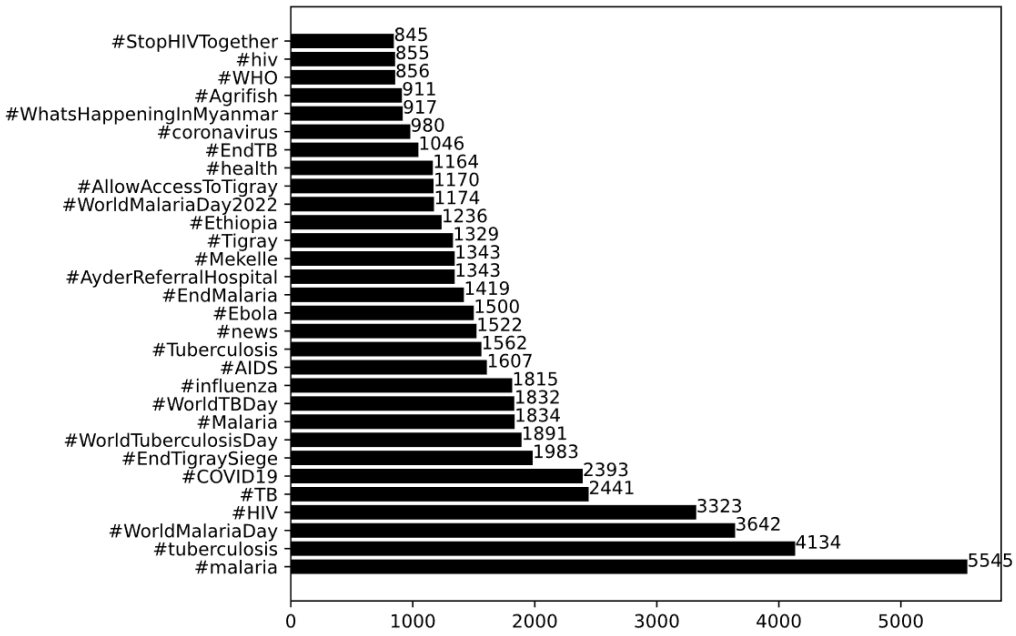


Figure 6. Top 30 hashtags

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

Apart from hashtags related to specific viruses, there are also other kind 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 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 that appeal Tigray needs urgent assistance to save lives.

As for #AyderReferralHospital due to limited medical resources, there are several patients infected with 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 turned away.

Figure 7 shows the number of tweets over time for different virus subsets. There are approximately 1,875 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 7 shows several peaks that can be put in relation with specific events in the real life: e.g. March 24th was the World Tuberculosis Day and April 25th was the World Malaria Day. The peak for HIV/AIDS is a little higher, i.e. 1,034, on March 10th, 2022, because there was the National Women and Girls HIV/AIDS Awareness Day. The peak of Tuberculosis tweets is 1,238 on April 29, 2022. Furthermore, April 24-30, 2022 week was the World Immunization Week 2022, when people's discussion about virus increase for Tuberculosis.

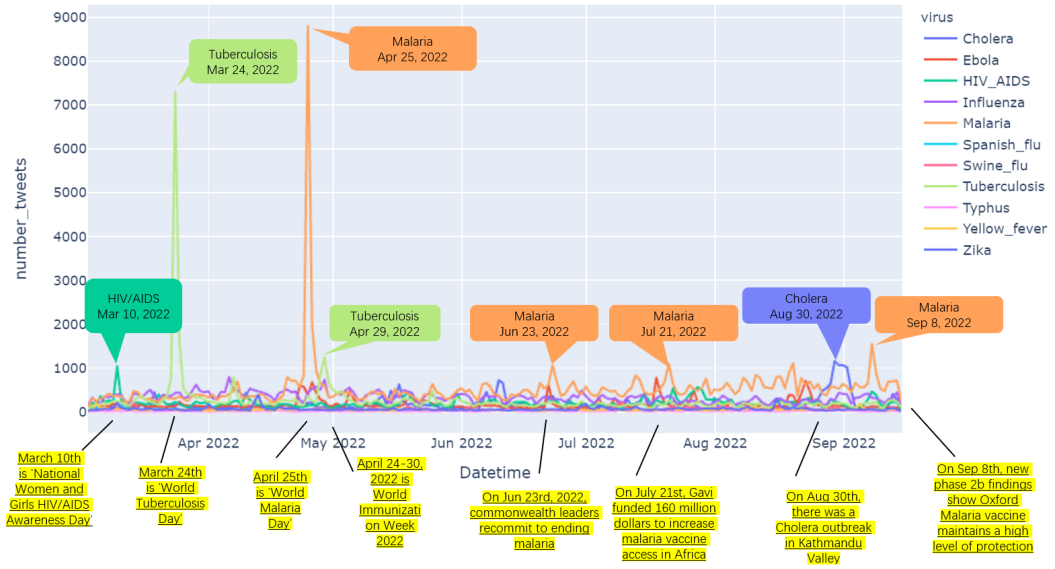


Figure 7. The number of tweets over time for the various epidemics

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

Figure 8 shows an earth map with the location of some tweets by considering the attribute place id, which is a unique identifier (ID) for location on Twitter. With this ID, we have been able to get detailed information about the place type, the full name of this place, and the country to which it belongs.

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 just obtains 6,863 tweets located.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

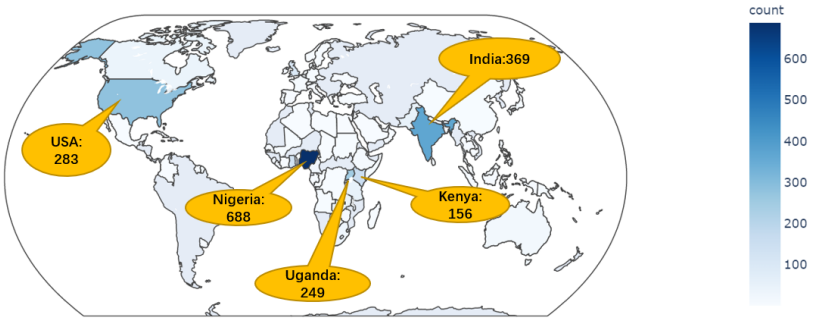


Figure 8. The location of tweets related to Malaria

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 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 9 shows a distribution of the breath symptom among the 11 viruses. To identify which symptoms account the highest percentage in a given dataset, we have calculated their frequency and its 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 weakness illness.

Different viruses have 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.

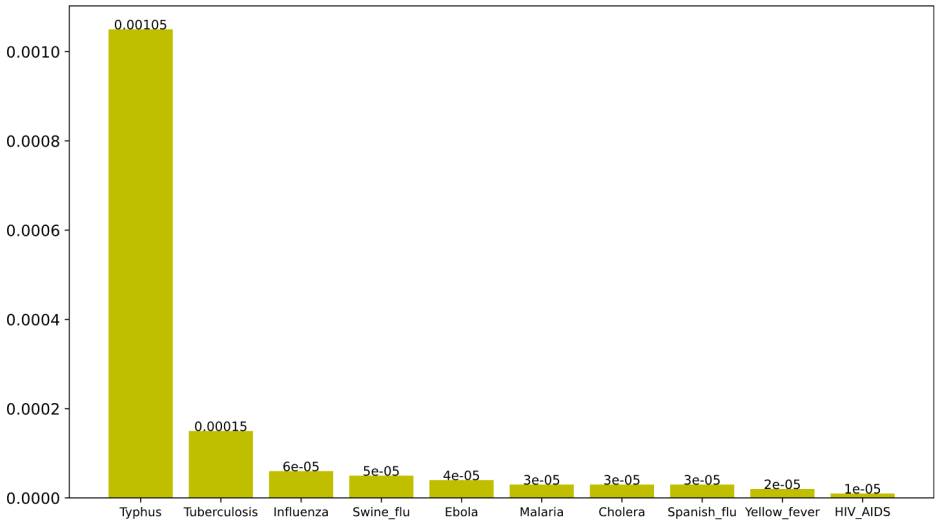


Figure 9. Dictionary and bar plot of breath problem

For breath problem (see Figure 9), tweets related to typhus are more often than other viruses. Similarly, compared to other viruses, discussion about fever accounts for a higher proportion in yellow fever dataset; compared with others, rash and diarrhea are typical symptoms for typhus and cholera respectively; cough is also widely discussed in tuberculosis dataset; death has a higher proportion in Spanish flu; according to history,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

there are exactly so much dead cases because of Spanish flu; jaundice is the feature of yellow fever; for other symptoms, there is no clear distinction among virus datasets.

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

Figure 10 shows retweets, replies and likes frequencies for the 11 pandemics. The number of retweets, replies and likes can be regarded as a symbol of interaction in social media. In general, the number of likes is much more than retweets and replies. Specifically, malaria has the most likes, replies and retweets. It means that tweets about malaria are popular on Twitter.

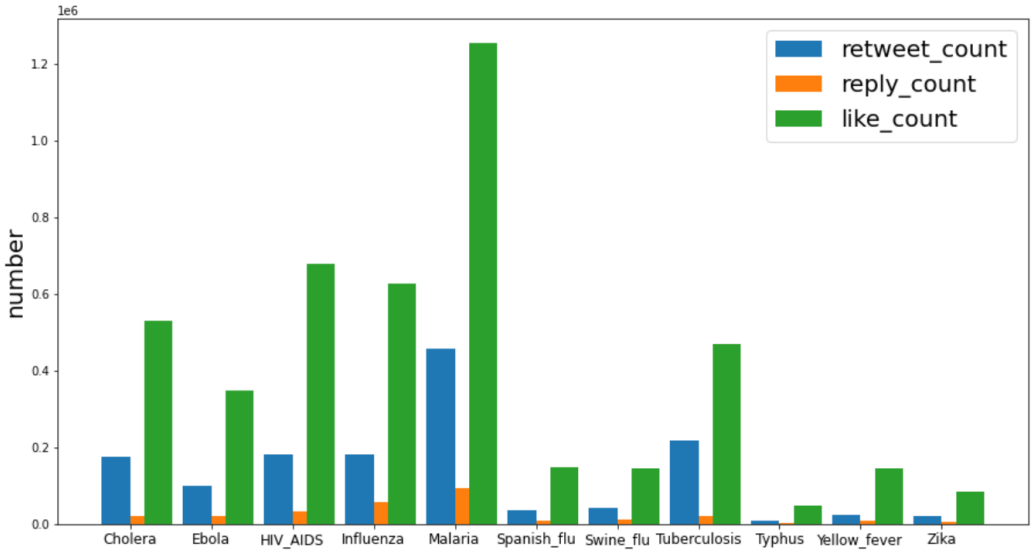


Figure 10. Retweets, replies and likes

5. Vectorization

In this study, four vectorizations have been considered: bag of words (BOW), 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 optimal size of word vector, in which the most important step is to evaluate the performance of word embedding. For example, Yin and Shen [42] introduce a Pairwise Inner Product loss function. In our study, we have considered Faruqui and Dyer [43]'s method that evaluates the performance of word embedding by calculating Spearman correlation between similarity score (regarded as ground truth) and cosine similarity in 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 maximizes the correlation. Figure 11 shows that 170 and 110 are the best numbers of dimension for Word2Vec and FastText respectively.

After ensuring the dimensionality and obtaining word embeddings, the interpretation of 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 4,199 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 maximizes 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 mean vector of all words within a sentence. Of course, there is a drawback that we may get same interpretation for different columns because of the size of the linguistic property.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

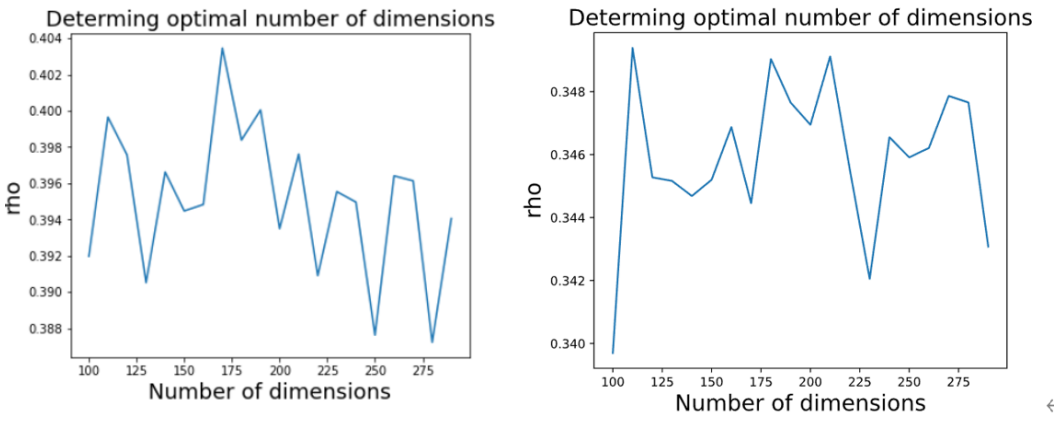


Figure 11. Dimension for Word2Vec on the left and FastText on the right

6. Clustering

In this study we have used *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. silhouette score [45]) from the cluster centroids for every *k*-means model: this value decreases when the number of clusters increases. Figure 12 shows that 8 is the best *k* choice for Word2Vec and FastText.

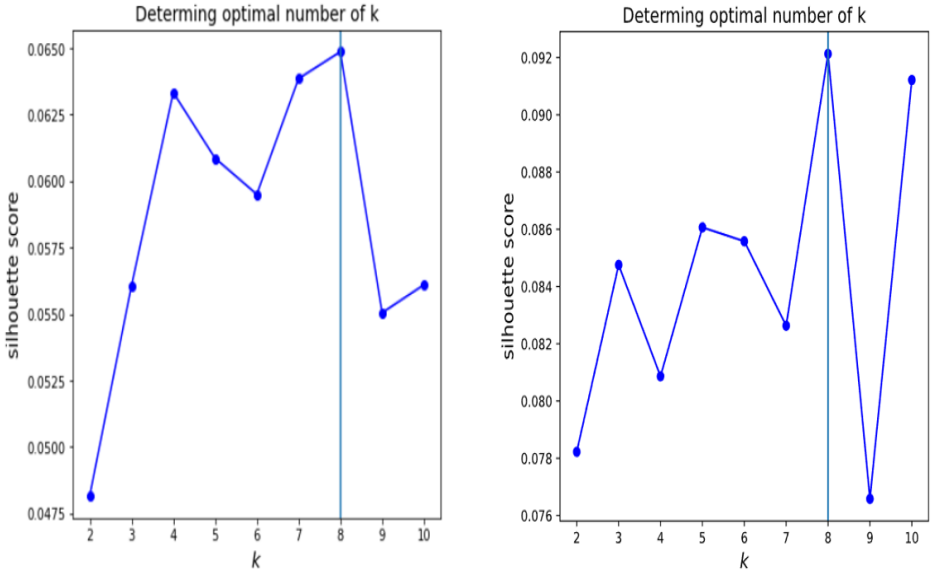


Figure 12. The choice of number of clusters - Word2VeC on the left side and FastText on the right side

In order to interpret the results, the top 30 words are listed according to their frequency in the different clusters. Table 1 and 2 summarize clusters based on Word2Vec and FastText respectively. To understand the similarity between 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

Cluster Number	Interpretation	Top words
cluster0	medical treatment	vaccine, disease, case
cluster1	people's health situation	old, baby, people
cluster2	cause of virus	mosquito, world
cluster3	war and conflict	Ukraine, war, refugee
cluster4	deaths	people, die, kill
cluster5	avian	bird, county, flock
cluster6	pandemic outbreak in Congo	Congo, outbreak, crisis
cluster7	people's activity like National Women and Girls HIV/AIDS Awareness Day	awareness, world

Table 1. Interpretation of clustering on Word2Vec

Cluster Number	Interpretation	Top words
cluster0	people's activity like National Women and Girls HIV/AIDS Awareness Day	people. awareness, national
cluster1	medical treatment	vaccine, peopel, mask
cluster2	fight against viruses	world, health , fight
cluster3	pandemic situation in a certain area	outbreak, Congo, city
cluster4	avian	bird, county, flock
cluster5	deaths	die, people, time
cluster6	war and conflict	Ukraine, Russia, kill
cluster7	cause of viruses	mosquito, world

Table 2. Interpretation of clustering on FastText

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

7. Topic Modeling

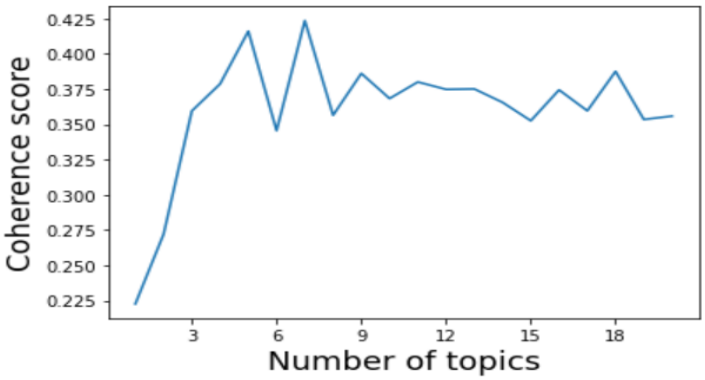
In our study we have applied the LDA model that performs several times the following operations to create documents: first, it selects one of the predefined topics with a certain probability, then selects a word under that topic with a certain probability. Assuming there are M documents with K topics. Each document (length N) has its own topic distribution that is polynomial with the parameters of the polynomial distribution that obeys the Dirichlet distribution and are α. Each topic has its own word distribution that is a multinomial distribution with the parameters of the multinomial distribution that obeys the Dirichlet distribution and are β. For the creation of 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 LDA. Apart from the input of sentence vector (Bag of words) and the dictionary (id and word), it is essential to specify the number of topics. In order to find optimal values, topic coherence is employed as the indicator to measure the performance of the model. It is meaningful to calculate the frequency of co-occurrence of 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 is varying from 1 to 20 to find an optimal value that maximizes the c_v coherence.

Figure 13 shows that 7 is the optimal value of the number of topics. The model with 7 topics can obtain a relatively high coherence score.

The LDA model can be used to extract the latent topics. Its results can be visualized with the pyLDAvis library [48] that is an open source package in Python to interactively present the result of LDA. Figure 14 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; Topic7 - other diseases like cancer. This result is too general because there are 11 different epidemics' sources.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1



h

Figure 13. Determining optimal number of topics

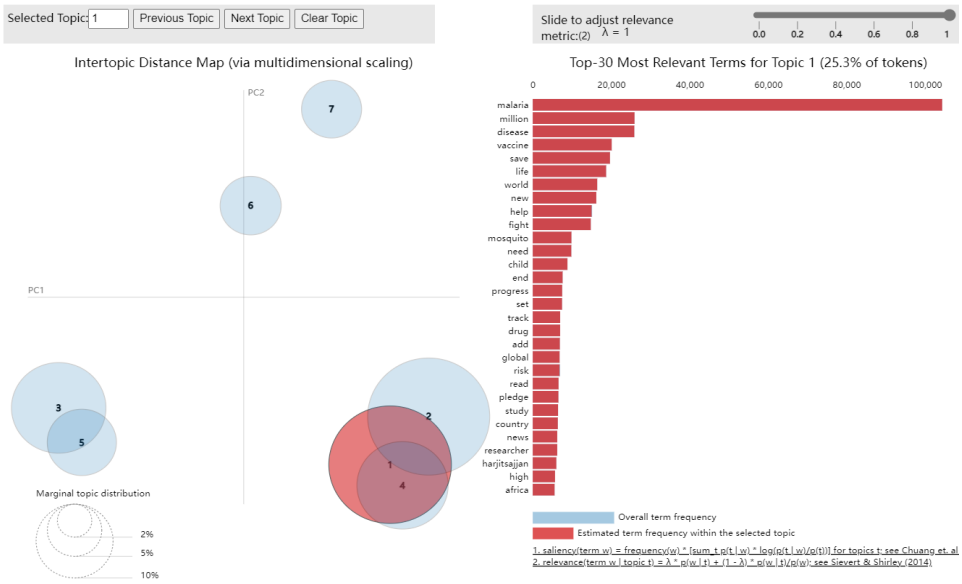


Figure 14. Topic modeling results based on LDA

8. Sentiment Analysis

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

There is no emotion or sentiment labels in our data, therefore, we have used 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 sentiments: for every tweets and corresponding emotions 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 label sentiments we have performed a semi-supervised learning. The dataset, as shown in Figure 15, has been divided into two parts: training data set (80%) and test data set (20%). Training data have been used to build classifier and testing data are used to measure the performance. Particularly, for training data, 80% of the labels of the data are got rid of and the remaining labels are regarded as the ground truth. We have defined classifiers by considering labeled training data, and using Logistic Regression (LR), Multinomial Bayes (MNB) model and Random Forest (RF) based on TF-IDF vector. Then, we have measured

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

their performances on our test data set, which are summarised in Table 3. According to the accuracy value of 0.80, we have selected the logistic regression model.

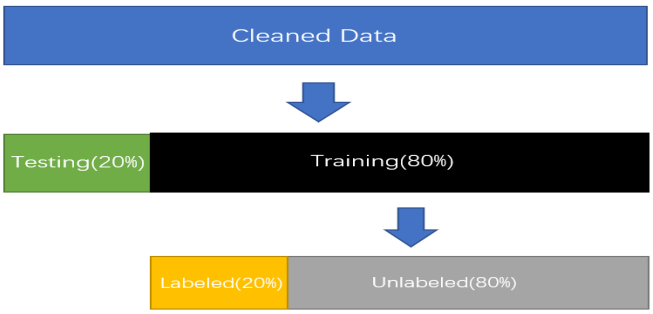


Figure 15. Splitting tweets data

Emotion	Support	LR			MNB			RF		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
anger	1,239	0.95	0.31	0.46	1.00	0.04	0.08	0.90	0.29	0.44
anticipation	4,496	0.83	0.61	0.70	0.99	0.07	0.13	0.79	0.45	0.58
disgust	9,659	0.81	0.86	0.84	0.93	0.29	0.45	0.76	0.83	0.80
fear	21,019	0.81	0.92	0.86	0.40	1.00	0.57	0.73	0.90	0.81
joy	5,263	0.85	0.67	0.75	0.99	0.15	0.27	0.82	0.52	0.63
neutral	3,115	0.77	0.74	0.76	1.00	0.07	0.14	0.59	0.93	0.72
sadness	5,391	0.86	0.64	0.73	0.99	0.14	0.24	0.85	0.53	0.50
surprise	778	0.93	0.32	0.48	1.00	0.01	0.01	0.90	0.35	0.50
trust	11,303	0.73	0.84	0.78	0.83	0.33	0.48	0.72	0.69	0.70
Index	Support	LR			MNB			RF		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
accuracy	62,263			**0.80**			0.48			0.74
macro avg	62,263	0.84	0.66	0.71	0.90	0.23	0.26	0.78	0.61	0.65
weighted avg	62,263	0.81	0.80	0.79	0.75	0.48	0.41	0.75	0.74	0.73

Table 3. Performance of Logistic Regression (LR), Multinomial Bayes model (MNB) and Random Forest (RF) with respect to emotion labels and indexes

We have also applied a self-training approach that belongs to the semi-supervised machine learning algorithms, as it uses a combination of labeled and unlabeled data to train the model. The idea behind the self-training approach consists of:

- using the labeled data to train the first supervised model, such as the logistic regression one;
- using the model to predict the class of unlabeled 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 labeled data;
- using the labels and pseudo-labels to train a new supervised model;
- making predictions again and adding the new observations to the pseudo-labeled pool;
- iterating these steps until no other unlabeled observations satisfy the pseudo labeling 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 sentiment of all observations.

Figure 16 shows that fear, trust and disgust are the three most dominant emotions. In detail, 38.4% of tweets have the fear emotion to pandemics and 16.5% of tweets contain the disgust mood. At the same time, there are 20.5% of tweets which express a trust emotion.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2022 doi:10.20944/preprints202211.0005.v1

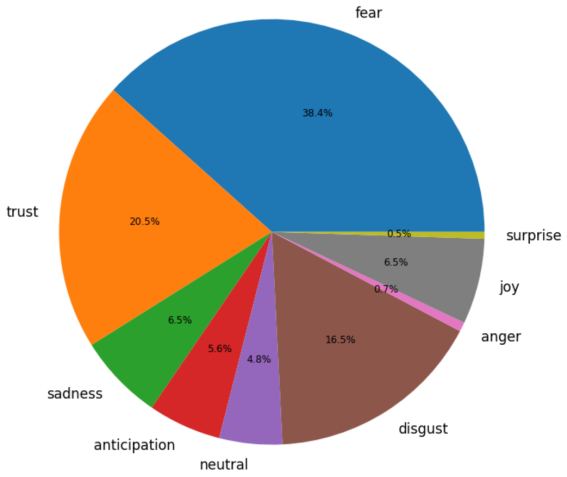


Figure 16. Emotion distribution

Figure 17 shows the sentiment distribution in different pandemic datasets: in cholera, spanish flu and yellow fever, fear dominates among 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; in malaria and swine flu, disgust has the biggest 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, like for HIV/AIDS, where people show a trust emotion.

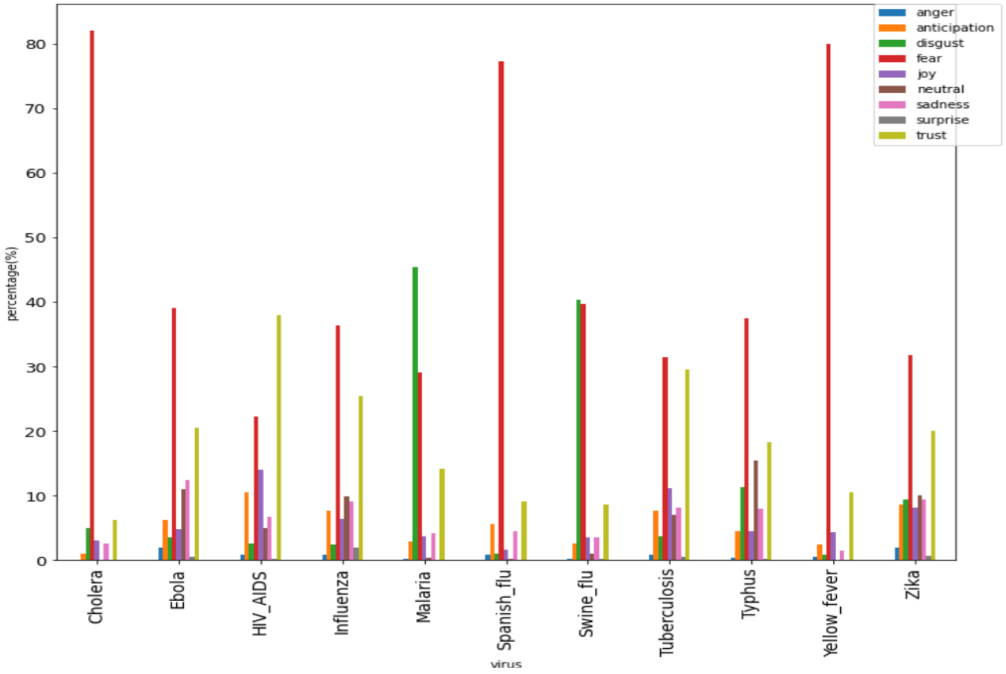


Figure 17. Emotion distribution in different subsets

9. Combining Topic Modelling and Sentiment Analysis

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

 doi:10.20944/preprints202211.0005.v1

Pandemic	Dominant Emotion	Topics
Ebola	fear	Topic1 - COVID-19
		Topic2 - Information about the virus
		Topic3 - Ebola outbreak in Congo
		Topic4 - Political issues in Congo like scandal
HIV / AIDS	trust	Topic1 - Research findings
		Topic2 - National Women and Girls HIV/AIDS Awareness Day
		Topic3 - High-risk or susceptible groups
		Topic4 - Help and encourage from experts
Malaria	disgust	Topic1 - Treatment like vaccine
		Topic2 - People's reaction
		Topic3 - The spreading of Malaria: the bite of mosquito
		Topic4 - Research about Malaria
		Topic5 - The cause of Malaria: the falciparum parasite
		Topic6 - Organisations or figures
		Topic7 - Usage of drug
		Topic8 - Medical treatment's achievement
Cholera	fear	Topic1 - Vaccine
		Topic2 - The risk of Cholera outbreak in Mariupol, Ukraine
		Topic3 - The cause of Cholera: water pollution
Influenza	fear	Topic1 - COVID-19
		Topic2 - Avian
		Topic3 - The risk of the virus
		Topic4 - New cases and deaths
Spanish flu	fear	Topic1 - COVID-19
		Topic2 - Avian
		Topic3 - Million deaths in the history
		Topic4 - Protective measures such as wearing masks
Swine flu	disgust	Topic1 - COVID-19
		Topic2 - The connection with avian influenza
		Topic3 - Influence on the world
		Topic4 - Animals like swine, monkey, and so on

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.

Tables 4 and 5 show that the fear emotion is dominant for 8 over 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 reasons of the virus, such as the bite of mosquito for malaria and the water pollution for cholera.

 doi:10.20944/preprints202211.0005.v1

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

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.

10. Discussion and Conclusions

In this work, we have used different natural language processing and machine learning techniques to explore pandemics' information in the 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 in the social media? What is the trend of these discussions over time?

Despite Covid-19 has been excluded from our analysis, the collected data show that Covid-19 is accounted by people discussion according to the frequency of hashtags and topic modeling. Furthermore, discussion about malaria, influenza and tuberculosis are the most in 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. For 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 Wpanish flu and swine flu; 3. nowadays typhus actually is 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 vectorization of each tweet message, we have computed the distance between vectors to explore observations' similarity by using k-means and then interpreted every cluster. Finally, according to topic modeling, 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 sentiment analysis, fear, trust and disgust are the three most dominant emotions. Specifically, we have presented the emotion distribution of every pandemic.

RQ4. What 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 pandemics. The frightening topics are often related to the cause of pandemics and their influence on human beings and society. It is worth highlighting that people have a fear emotion to several medical treatments like wearing masks or taking vaccine although these measures are effective in controlling pandemics. In order to eliminate people's fear, it is

 doi:10.20944/preprints202211.0005.v1

important for 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 influence of nature, there are some much human factors in the outbreak of these pandemics. Our findings show that there are also positive emotions or attitudes. For example, people have trust emotion in tweets related to HIV/AIDS.

Appendix A. Description of the Considered Pandemics

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

Ebola is a viral hemorrhagic 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 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.

The **Spanish influenza**, or 1918 flu pandemic, was an unusual deadly influenza pandemic that broke out between January 1918 and April 1920. The main symptoms are sore throat, headache, fever and mucosal hemorrhage; but it leads to death if it becomes severe.

The **swine flu** is an infection caused by several types of swine influenza viruses. Swine influenza virus is common throughout pig populations. If its transmission causes human flu, it is called zoonotic swine flu. Symptoms of 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 microbacterium tuberculosis bacteria that mainly affects the lungs. Its symptom is 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 characterized by severe fever and jaundice. It is caused by 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.

Author Contributions: For Conceptualization, 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.; visualization, 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.

 doi:10.20944/preprints202211.0005.v1

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

References

1. Morens, D.M.; Folkers, G.K.; Fauci, A.S. The challenge of emerging and re-emerging infectious diseases. *Nature* **2004**, *430*, 242–249. https://doi.org/10.1038/nature02759.
2. Fan, V.; Jamison, D.; Summers, L. The Inclusive Cost of Pandemic Influenza Risk. Technical report, 2016. https://doi.org/10.3386/w22137.
3. III, F.J.G.; Sheps, S.; Ho, K.; Novak-Lauscher, H.; Eysenbach, G. Social Media: A Review and Tutorial of Applications in Medicine and Health Care. *Journal of Medical Internet Research* **2014**, *16*, e13. https://doi.org/10.2196/jmir.2912.
4. PAUL, M.J.; SARKER, A.; BROWNSTEIN, J.S.; NIKFARJAM, A.; SCOTCH, M.; SMITH, K.L.; GONZALEZ, G. SOCIAL MEDIA MINING FOR PUBLIC HEALTH MONITORING AND SURVEILLANCE. In Proceedings of the Biocomputing 2016. WORLD SCIENTIFIC, 2015. https://doi.org/10.1142/9789814749411_0043.
5. Vilic, A.; Petersen, J.A.; Hoppe, K.; Sorensen, H.B.D. Visualizing patient journals by combining vital signs monitoring and natural language processing. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016. https://doi.org/10.1109/embc.2016.7591245.
6. Tissot, H.C.; Shah, A.D.; Brealey, D.; Harris, S.; Agbakoba, R.; Folarin, A.; Romao, L.; Roguski, L.; Dobson, R.; Asselbergs, F.W. Natural Language Processing for Mimicking Clinical Trial Recruitment in Critical Care: A Semi-Automated Simulation Based on the LeoPARDS Trial. *IEEE Journal of Biomedical and Health Informatics* **2020**, *24*, 2950–2959. https://doi.org/10.1109/jbhi.2020.2977925.
7. Zhang, X.; Saleh, H.; Younis, E.M.G.; Sahal, R.; Ali, A.A. Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System. *Complexity* **2020**, *2020*, 1–10. https://doi.org/10.1155/2020/6688912.
8. Sepúlveda, A.; Periñán-Pascual, C.; Muñoz, A.; Martínez-España, R.; Hernández-Orallo, E.; Cecilia, J.M. COVIDSensing: Social Sensing Strategy for the Management of the COVID-19 Crisis. *Electronics* **2021**, *10*, 3157. https://doi.org/10.3390/electronics10243157.
9. Imran, M.; Qazi, U.; Ofli, F. TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels. *Data* **2022**, *7*, 8. https://doi.org/10.3390/data7010008.
10. Graff, M.; Moctezuma, D.; Miranda-Jiménez, S.; Tellez, E.S. A Python library for exploratory data analysis on twitter data based on tokens and aggregated origin–destination information. *Computers &amp Geosciences* **2022**, *159*, 105012. https://doi.org/10.1016/j.cageo.2021.105012.
11. Cornelius, J.; Ellendorff, T.; Furrer, L.; Rinaldi, F. COVID-19 Twitter Monitor: Aggregating and Visualizing COVID-19 Related Trends in Social Media. In Proceedings of the Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task; Association for Computational Linguistics: Barcelona, Spain (Online), 2020; pp. 1–10.
12. Andreadis, S.; Antzoulatos, G.; Mavropoulos, T.; Giannakeris, P.; Tzionis, G.; Pantelidis, N.; Ioannidis, K.; Karakostas, A.; Gialampoukidis, I.; Vrochidis, S.; et al. A social media analytics platform visualising the spread of COVID-19 in Italy via exploitation of automatically geotagged tweets. *Online Social Networks and Media* **2021**, *23*, 100134. https://doi.org/10.1016/j.osnem.2021.100134.
13. Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C.M.; Brugnoli, E.; Schmidt, A.L.; Zola, P.; Zollo, F.; Scala, A. The COVID-19 social media infodemic. *Scientific Reports* **2020**, *10*. https://doi.org/10.1038/s41598-020-73510-5.
14. Biancovilli, P.; Makszin, L.; Jurberg, C. Misinformation on social networks during the novel coronavirus pandemic: a quali-quantitative case study of Brazil. *BMC Public Health* **2021**, *21*. https://doi.org/10.1186/s12889-021-11165-1.
15. Househ, M. Communicating Ebola through social media and electronic news media outlets: A cross-sectional study. *Health Informatics Journal* **2016**, *22*, 470–478. https://doi.org/10.1177/1460458214568037.
16. Yousefinaghani, S.; Dara, R.; Poljak, Z.; Bernardo, T.M.; Sharif, S. The Assessment of Twitter's Potential for Outbreak Detection: Avian Influenza Case Study. *Scientific Reports* **2019**, *9*. https://doi.org/10.1038/s41598-019-54388-4.
17. Aramaki, E.; Maskawa, S.; Morita, M. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In Proceedings of the Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Edinburgh, Scotland, UK., 2011; pp. 1568–1576.
18. Santillana, M.; Nguyen, A.T.; Dredze, M.; Paul, M.J.; Nsoesie, E.O.; Brownstein, J.S. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology* **2015**, *11*, e1004513. https://doi.org/10.1371/journal.pcbi.1004513.
19. Gori, D.; Reno, C.; Remondini, D.; Durazzi, F.; Fantini, M.P. Are We Ready for the Arrival of the New COVID-19 Vaccinations? Great Promises and Unknown Challenges Still to Come. *Vaccines* **2021**, *9*, 173. https://doi.org/10.3390/vaccines9020173.
20. Sicilia, R.; Giudice, S.L.; Pei, Y.; Pechenizkiy, M.; Soda, P. Twitter rumour detection in the health domain. *Expert Systems with Applications* **2018**, *110*, 33–40. https://doi.org/10.1016/j.eswa.2018.05.019.
21. Durazzi, F.; Müller, M.; Salathé, M.; Remondini, D. Clusters of science and health related Twitter users become more isolated during the COVID-19 pandemic. *Scientific Reports* **2021**, *11*. https://doi.org/10.1038/s41598-021-99301-0.
22. Mahdikhani, M. Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic. *International Journal of Information Management Data Insights* **2022**, *2*, 100053. https://doi.org/10.1016/j.jjimei.2021.100053.

23. Bellandi, V.; Ceravolo, P.; Maghool, S.; Siccardi, S. A Comparative Study of Clustering Techniques Applied on Covid-19 Scientific Literature. In Proceedings of the 2020 7th International Conference on Internet of Things: Systems, Management and Security (IOTSMS). IEEE, 2020. https://doi.org/10.1109/iotsms52051.2020.9340213.
24. Alsudias, L.; Rayson, P. Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study. *JMIR Medical Informatics* **2021**, *9*, e27670. https://doi.org/10.2196/27670.
25. tweepy. Tweepy Documentation. https://docs.tweepy.org/en/stable/, Looked at October 16, 2022.
26. spacy. Industrial-Strength Natural Language Processing in Python. https://spacy.io/, Looked at October 16, 2022.
27. NLTK. NLTK Documentation. https://www.nltk.org/$_{m}$*odules*/*nltk*/*stem*/*wordnet.html*, *LookedatOctober*16, 2022.
28. pypi. autocorrect 2.6.1. https://pypi.org/project/autocorrect/, Looked at October 16, 2022.
29. Karthika, P.; Murugeswari, R.; Manoranjithem, R. Sentiment Analysis of Social Media Network Using Random Forest Algorithm. In Proceedings of the 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). IEEE, 2019. https://doi.org/10.1109/incos45849.2019.8951367.
30. Alodadi, M.; Janeja, V.P. Similarity in Patient Support Forums Using TF-IDF and Cosine Similarity Metrics. In Proceedings of the 2015 International Conference on Healthcare Informatics. IEEE, 2015. https://doi.org/10.1109/ichi.2015.99.
31. Jacobson, O.; Dalianis, H. Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. In Proceedings of the Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2016. https://doi.org/10.18653/v1/w16-2926.
32. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. FastText.zip: Compressing text classification models **2016**. [arXiv:cs.CL/1612.03651].
33. Kappus, P.; Groß, P. Finding Clusters of Similar-minded People on Twitter Regarding the Covid-19 Pandemic **2022**. [arXiv:cs.SI/2203.04764]. https://doi.org/10.5121/csit.2021.111803.
34. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
35. Qorib, M.; Oladunni, T.; Denis, M.; Ososanya, E.; Cotae, P. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications* **2023**, *212*, 118715. https://doi.org/10.1016/j.eswa.2022.118715.
36. WHO. Ebola virus disease - Democratic Republic of the Congo. https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON377, Looked at October 16, 2022.
37. BBC. Cholera in Mariupol: Ruined city at risk of major cholera outbreak - UK. https://www.bbc.com/news/wo europe-61762787, Looked at October 16, 2022.
38. Wikipidia. Queensland tick typhus. https://en.wikipedia.org/wiki/Queensland$_{t}$*ick*$_{t}$*yphus*, *LookedatOctober*16
39. KMH. Yellow fever - Kenya. https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON361, Looked at October 16, 2022.
40. UN. Ethiopia: Essential aid reaches Tigray region, but more still needed. https://news.un.org/en/story/2022/0 Looked at October 16, 2022.
41. Telegraph, T. Let's die at home: 200 patients turned away as Tigray's main hospital runs out of supplies. https://www.telegraph.co.uk/global-health/terror-and-security/die-home-200-patients-turned-away-tigrays-main-hospital-runs/, Looked at October 16, 2022.
42. Yin, Z.; Shen, Y. On the Dimensionality of Word Embedding, 2018. https://doi.org/10.48550/ARXIV.1812.04224.
43. Faruqui, M.; Dyer, C. Community evaluation and exchange of word vectors at wordvectors. org. In Proceedings of the Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 19–24.
44. Tsvetkov, Y.; Faruqui, M.; Ling, W.; Lample, G.; Dyer, C. Evaluation of word vector representations by subspace alignment. In Proceedings of the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2049–2054.
45. Shahapure, K.R.; Nicholas, C. Cluster Quality Analysis Using Silhouette Score. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2020. https://doi.org/10.1109/dsaa49011.2020.00096.
46. Hubert, L.; Arabie, P. Comparing partitions. *Journal of Classification* **1985**, *2*, 193–218. https://doi.org/10.1007/bf01908075.
47. gensim. gensim 4.2.0. https://pypi.org/project/gensim/, Looked at October 16, 2022.
48. Sievert, C.; Shirley, K. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Association for Computational Linguistics, 2014. https://doi.org/10.3115/v1/w14-3110.

 doi:10.20944/preprints202211.0005.v1

49. Mohammad, S.M. Word Affect Intensities. In Proceedings of the Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018); , 2018.
50. NRC. emotion-nrc-affect-lex 0.0.3. https://pypi.org/project/emotion-nrc-affect-lex/, Looked at October 16, 2022.
51. CDC. Epidemic Typhus. https://www.cdc.gov/typhus/epidemic/index.html, Looked at October 11, 2022.