Covid-19 Twitter Data Analysis Using Natural Language Processing

Dželila Mehanović[*1], Zerina Mašetić[1], Amela Vatreš[1]

[1]Faculty of Engineering and Natural Sciences, International Burch University

*Corr. author email: dzelila.mehanovic@ibu.edu.ba
ORCID No. 0000-0001-7731-0478, 0000-0002-6226-8868

Abstract:

The world and everyday life are completely changed due to novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) which caused COVID-19 disease. New decisions and rules are set from day to day. Reactions and attitudes of people are often shared on social media, especially using Twitter which makes Twitter a good source of data for many researches. In this work, we analyse a dataset of tweets that have at least 1000 retweets to perform sentiment analysis. Python libraries were used for analysis which includes language detection and translation, preprocessing, tokenization and stemming. Moreover, sentiment analysis and classification is performed using Naive Bayes. It is found that most of the tweets have positive sentiment and achieved accuracy is 60%. Classifier is able to predict positive and neutral tweets with high rate, while it has lower accuracy rate for negative sentiment.

Keywords: twitter, covid-19, natural language processing, sentiment analysis, classification, Naive Bayes

1. Introduction

The first confirmed cases of Covid-19 appeared in December 2019 in China, but it spread all over the world very fastly. There are more than 23 million confirmed cases in the world [1]. The virus caused huge problems and made changes in the people's lives. The way how education, healthcare and finance systems operate, the way how people work and interact were affected and changed completely to adjust to the new situation [2].

People have different opinions regarding the virus. There are those who do not believe that the virus exists, and those who believe that it exists and is dangerous to the humans. Also, people tend to share their opinions and feelings over social

networks. Twitter is social media that allows people to share their thoughts in form of tweets which can be long up to 280 characters. These tweets are used as a source of data for many researchers since it provides insight in the state of mind of people.

According to the statistics there are 500 million tweets every day which means that around 6000 tweets are shared in one second. Also, there are 145 million active users that visit and exchange tweets every day [3]. Number of tweets and users make twitter a great data source for many different research topics [4]. Twitter API is available for use to retrieve and manipulate twitter data [5].

Natural language processing is used for a long time to manipulate natural language data and provide outputs useful to build different applications that make life easier. Sentiment analysis is one of the natural language processing methods that is used to decide whether data is positive, negative or neutral. Sentiment analysis can help to follow opinions, needs and problems of people. Also, it helps to identify most common topics discussed in public.

In this work, we apply natural language processing techniques to analyse twitter data, perform sentiment analysis and implement classification of tweets. The rest of the work is organized as follows: in section two we provide literature review, that is we give overview of related works and description of natural language processing, text analysis, classification, and twitter data. In section three we describe methodologies and steps that are performed in the experiment. Section four contains results of the experiment described in section three. Finally, with section five we conclude our work.

2. Literature Review

In this section we present details about works that are related to our topic. After that natural language processing, data analysis, classification and classification methods commonly used with natural processing and twitter data are described.

2.1. Related Works

Authors in [6] public sentiment about covid-19 using twitter data. They tried a two classifiers Naive Bayes which achieved accuracy 91% and logistic regression whose performance was 74%. [7] performed analysis of twitter data with hashtags coronavirus and covid-19 from 9th April to 15th April 2020. After analysis they found that more than 50% were neutral and a large portion of tweets were objective. But also, they found that public sentiment varies from day to day. [8] searched for twitter data with terms: “corona”, “2019-nCov,” and “COVID-19”. Tweets are analyzed using word frequencies of unigrams and bigrams and using latent Dirichlet allocation tweets are categorized in four topics. Ten of the topics had positive and

2 of them had negative sentiment. The highest mean of likes was 15.4 for tweets related to economic loss and lowest was 3.94 for tweets about travel bans and warnings. [9] analyzed twitter data from March 20 to April 19, 2020 using sentiment analysis 48.2% tweets are classified as positive, 20.7% as neutral and 31.1% as negative. Also, five topics are identified as the most common health care environment, emotional support, business economy, social change, and psychological stress. After quarantine started, [10] processed twitter data from the Philippines using natural language processing and sentiment analysis to determine effects of quarantine to lifestyle. It is determined that most users have negative sentiments and the most problems are related to the food supply and support from the government. They identified that negative sentiment rises over time and it is expected to continue.

2.1. Natural Language Processing

Term natural language denotes language that is used in standard communication between people. Natural language processing (NLP) refers to the manipulation of natural languages performed by computers [11]. It includes various text analysis such as word frequency analysis, sentiment analysis, language translation or giving response to some request.

NLP became widely used in the last two decades, so there are many real word applications that utilize natural language processing such as smart assistants like Siri, Cortana and Alexa, autocorrect and autocomplete functionalities, email classification, chatbots [12] [13]. SignAll application [14] uses natural language processing to convert language to text, which enables deaf people to communicate with those who do not know sign language. Well known Google Translate application [14], [15] uses natural language processing to translate text.

2.2. Text Analysis

Text analysis is a process of text processing to extract useful information [16]. There are different types of methods that can be used in text analysis from basic methods such as word frequency, collocation, concordance to those more advanced such as text classification which includes sentiment analysis, keyword extraction and topic detection.

Word frequency analysis is used to find words that are mostly used in the given context or conversation. Collocation reveals words that are commonly used together with. In the collocation, bigrams, trigrams or quadrans are identified. Concordance is used to find context of the words in the text or set of words. [1]

Classification is used to assign labels to the text. Sentiment analysis is used to detect emotions in a text such as positive, negative or neutral. Topic or category detection is another type of classification used in NLP. Text is classified as one of categories, for example news, sport, entertainment or music. Keyword extraction refers to extraction of the most important words from the text. It can be used to

summarize text, create indexes on data that is going to be searched or to create word clouds to provide visual representation of the text.[1]

2.3. Classification

Classification is a technique used to assign the correct label to the given input [1]. Label is selected from the list of predefined labels. Email filtering, text topic and word context detection are examples of classification tasks.
There are many machine learning classification algorithms that can be applied to perform classification tasks. Some of these classification algorithms are Naive Bayes, Decision Trees, Artificial Neural Network, K-Nearest Neighbor. Decision trees are used to generate tree charts which contain assigned labels to input based on input features. Naive Bayes is a classification algorithm in which each feature contributes to output independently [1]. It is called a probabilistic classifier since it selects the class that has the highest probability to be the correct class.

There are three datasets used in the process of classification: training, test and dev-test dataset. Training set is used to train the classifier, the test set is used to evaluate classifier performance and the dev-test set is used to adjust model features.

Accuracy is used to measure the performance of a classifier and represents the ratio of correctly classified items over total number of items [17].

3. Methodology

3.1. Data

We used a tweeter dataset downloaded from [18]. This dataset contains 48751 tweets related to the covid-19.

The following are several examples of tweets from dataset:

- It looks like the Democrats, along with the left-wing media, are going to try to use the Coronavirus to weaponize it against the 2020 Election Process!
- Every single person should be participating in Coronavirus response by taking the same simple precautions every day.
- While we cannot stop the spread of #coronavirus, we can SLOW it, decreasing the risk of overwhelming our hospitals.
- Another existing med shows great promise killing the virus!
- A few minutes ago, I just found out that my aunt who was a #healthcareworker in #NewYork has died from #COVID19. She contracted the virus from one of her patients. We had hoped she would recover. I am heartbroken and devastated.

Total number of generated tweets in the period between January and May is 628,809,016 which presents a large data set. For purposes of this study, we used a smaller version of the dataset. This version contains 43751 tweets that have at least 1000 retweets. These tweets are collected from January 22 to April 23. It is

important to mention that these tweets are considered as important due to the number of retweets, but those do not have to be the most important tweets related to the covid-19. Table 1 and Fig.1 shows distribution of tweets by period of time.

Table 1. Distribution of tweets in time

Date period	# of tweets
January 22 - March 11	5531
March 12 - March 21	15000
March 21 - April 7	15000
April 7 - April 13	4110
April 13 - April 23	2945

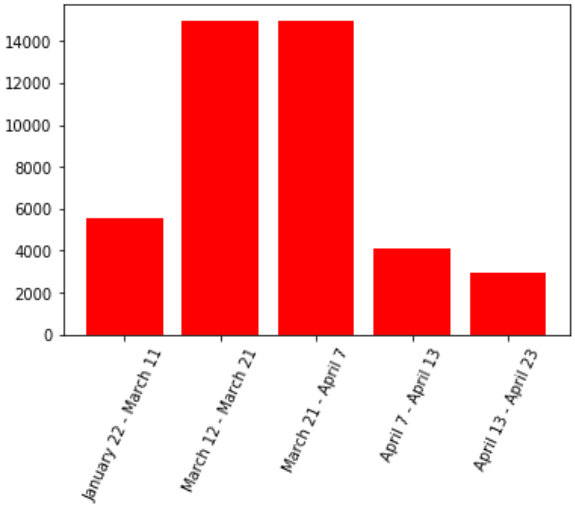

Fig. 1. Distribution of tweets in time

Dataset contains multiple features such as user, date, tweet, permalink, retweet count and likes count. For our research we used tweet data to perform sentiment analysis and classification.

3.2. Language detection and translation

Data set contains tweets in 43 different languages such as english, french, spanish, portuguese and others. To detect the language of tweets we used python library langdetect [19] text given as argument. It supports 55 languages and returns the ISO 639-1 Code.

After we detected language in which tweets are written, we used another python library to translate all tweets to the English language. Googletrans library [20] implements Google Translate API and it is used to translate text from one language to another.

3.3. Data preprocessing

Collected tweets contained mixed data that included urls, special characters and tags. To proceed with analysis, it was necessary to perform data preprocessing. Data preprocessing can include several stages such as data cleaning, integration, transformation, reduction and discretization [21]. To perform data preprocessing we mostly used python library re [22] for regex. HTML tags, punctuation marks, numbers, hashtags and urls are removed from tweet data. Also, all letters are transformed into lowercase letters. These steps done in preprocessing made creating a bag of words easier.

3.4. Word tokenization

Word tokenization is a process of separating text or sentences into words [23]. Words have to be separated and identified in order to be used with natural language processing. To tokenize words we used the word_tokenize method from the nltk tool available in python [24].

3.5. Stemming and stopwords

Words that have the same base are used in different formats within text, that is with different suffixes. Stemming is a process that is used to return words to their base format [25]. Before we applied stemming, more precisely Porter Stemmer [26] to our dataset, we firstly removed all stopwords. To remove stop words we used corpus of stopwords from python nltk [27].

3.6. Sentiment analysis

As described before, sentiment analysis is used to detect public sentiment about some topic. Here we want to find sentiment related to tweets about covid-19. Our dataset does not contain labeled data, so before classification we had to detect sentiment of each tweet.

Two packages were used to detect sentiment of tweets. Those are nltk sentiment analyzer [28] and textblob package [29]. By review of results we found that textblob provided more accurate sentiment analysis. After this phase, tweets are labeled as positive, neutral or negative.

3.7. Classification

To classify tweets, we used Naive Bayes classifier. Firstly, we build a list of tuples which contain tweet and sentiment value. Secondly, we build a list of all unique words which are checked to be English words. Using these words and a list of tweets we build three datasets of features: those with only positive, only neutral and only negative tweets. Finally, the dataset is divided into training and testing sets with ratio 75:25.

4. Results

In this section we present results of the experiment described in the previous section.

We managed to detect 43 different languages in which tweets were written. Table 2 presents the distribution of tweets per language.

Table 2. Tweets distribution per language

Language code	# of tweets	Language code	# of tweets	Language code	# of tweets
en	26331	ta	33	sw	3
es	9080	nl	17	hr	3
pt	3196	ur	16	sv	3
fr	1744	lt	12	mk	2
th	933	no	11	pl	2
id	689	et	7	fi	2
ca	372	af	7	zh-cn	2
ja	340	ro	5	hu	1
it	243	da	5	bn	1
hi	225	so	5	el	1
de	119	vi	5	zh-tw	1

tr	103	ru	4	te	1
tl	92	cy	4	cs	1
ko	41	sk	4		
ar	40	sl	4		

After detection of language, all tweets are translated into English language. Average tweet length and number of words are counted using translated values. Average tweet length was 253 characters and average number of words was 51. Furthermore, we created word clouds using english values. Word cloud is used to present most frequently used words. Several word clouds are created, below we provide two of them. Fig. 2 and Fig. 3 show two word clouds. It is noticeable that coronavirus, covid19, China, Trump and cases are words that appear in both of them and seem to be most frequently used.

Fig. 2. Word cloud 1

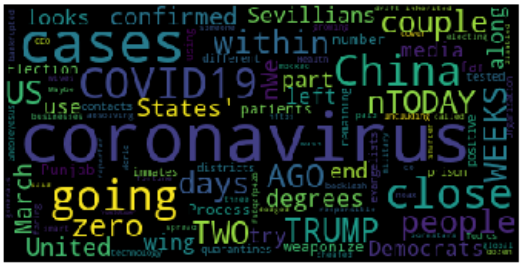

Fig. 3. Word cloud 2

Sentiment analysis was used to label tweets as positive, neutral or negative. Distribution of labels is presented in Table 3 and Fig. 4.

Table 3. Sentiment distribution

Sentiment	# of tweets
positive	20324
neutral	12582
negative	10845

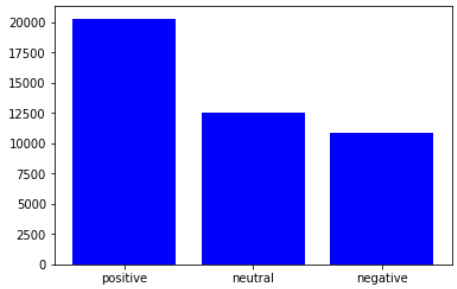

Fig. 4. Sentiment distribution

From the sentiment analysis, we found that most tweets are positive, and the number of negative tweets is two times lower than the number of positive tweets. Moreover, we analysed sentiment distribution of tweets per language. Table 4 presents a number of positive, negative and neutral tweets for 4 languages for which we have more than 1000 tweets. It can be noticed that for all of them the number of positive tweets is highest.

Table 4. Sentiment distribution per languages

Language	# of positive tweet	# of neutral tweets	# of negative tweets
en	12732	6793	6806
es	3986	2882	2212
pt	1328	1132	736
fr	679	669	396

For classification, we build list of tweet-sentiment value pairs that comes in the following form:

[(tweet. sentiment), (tweet2, sentiment), ...]

Next step was to build a list of all unique words. For this task, we used words that are given as output from the stemming phase. Finally, datasets for positive, negative and neutral features are obtained. All three datasets are combined in one which is further divided into training and test sets.

Classifier is built using a training set and evaluated using a test set. But, when we applied the entire training dataset, we experienced the problem with the memory. Because of that, classification is applied on a smaller dataset with randomly selected instances. To evaluate the performance, we used accuracy which was 60% where the classifier seemed to work better with positive and neutral tweets.

5. Conclusion

In this work we present analysis of tweets related to the covid-19. Dataset includes tweets that have at least 1000 retweets. We performed word frequency analysis and detected the language of tweets. In the language detection phase, we found that tweets are written in 43 different languages, and the python library is used to translate all tweets to English language to make further steps easier.

Tweets did not contain sentiment labels, so it was necessary to perform sentiment analysis to label those. Using two libraries, sentiment of tweets are discovered and a more reliable analyzer TextBlob is selected. It is found that most analysed tweets are positive and there are two times more positive than negative tweets.

Before sentiment analysis, tweets are preprocessed and word tokenization is performed. Moreover, we removed stop words and applied Porter Stemmer to perform stemming. Finally, we created datasets with features for positive, negative and neutral tweets. These datasets are combined to build a training and test set with a ratio 75%-25%.

Classification is performed using Naive Bayes classification algorithm and achieved accuracy is 60%. It is found that positive and neutral tweets are detected with higher accuracy rates, while accuracy for negative sentiment was lower.

This kind of analysis using natural language processing is useful to track public opinions, problems and reactions to decisions that are made and affect groups of people. To expand research, we could use a dataset with geographical location to have more specific insight to sentiment based on location. Also, a classifier should be improved to detect negative sentiment more precisely.

References

[1] Coronavirus Update (Live): 23,838,486 Cases and 817,636 Deaths from COVID-19 Virus Pandemic - Worldometer (no date). Available at: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdUOA?Si (Accessed: 25 August 2020).

[2] Badnjević, A. et al. (2020) 'Risks of emergency use authorizations for medical products during outbreak situations: a COVID-19 case study', Biomedical engineering online, 19(1), p. 75.

[3] Kim, A. E. et al. (2013) 'Methodological considerations in analyzing Twitter data', Journal of the National Cancer Institute. Monographs, 2013(47), pp. 140–146.

[4] Getting Started with the Twitter API (no date). Available at: https://developer.twitter.com/en/docs/twitter-api/getting-started/guide (Accessed: 25 August 2020).

[5] How many tweets about Covid-19 and Coronavirus? 508 MM tweets so far (2020). Available at: https://www.tweetbinder.com/blog/covid-19-coronavirus-twitter/ (Accessed: 24 August 2020).

[6] Samuel, J. *et al.* (no date) 'COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification'. doi: 10.31234/osf.io/sw2dn.

[7] Manguri, K. H., Ramadhan, R. N. and Mohammed Amin, P. R. (2020) 'Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks', Kur-distan Journal of Applied Research, pp. 54–65. doi: 10.24017/covid.8.

[8] Abd-Alrazaq, A. et al. (2020) 'Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study', Journal of medical Internet research, 22(4), p. e19016.

[9] Hung, M. et al. (2020) 'Social Network Analysis of COVID-19 Senti-ments: Application of Artificial Intelligence', Journal of medical Internet research, 22(8), p. e22590.

[10] 'SENTIMENT ANALYSIS OF FILIPINOS AND EFFECTS OF EXTREME COMMUNITY QUARANTINE DUE TO CORONAVIRUS (COVID-19) PANDEMIC' (2020) *Journal of critical reviews*. doi: 10.31838/jcr.07.07.15.

[11] Bird, S., Klein, E. and Loper, E. (2009) Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. 'O'Reilly Media, Inc.'

[12] *8 common examples of natural language processing and their impact on communication* (no date). Available at: https://www.tableau.com/learn/articles/natural-language-processing-examples (Accessed: 25 August 2020).

[13] Kharkovyna, O. (2019) Natural Language Processing (NLP): Top 10 Applications to Know, Towards Data Science. Available at: https://towardsdatascience.com/natural-language-processing-nlp-top-10-applications-to-know-b2c80bd428cb (Accessed: 25 August 2020).

[14] Coldewey, D. (2018) SignAll is slowly but surely building a sign language translation platform, TechCrunch. Available at: http://techcrunch.com/2018/02/14/signall-is-slowly-but-surely-building-a-sign-language-translation-platform/ (Accessed: 25 August 2020).
[15] Google Translate (no date). Available at: https://translate.google.com/ (Accessed: 25 August 2020).
[16] "*Text Analysis* (no date). Available at: https://monkeylearn.com/text-analysis (Accessed: 25 August 2020).
[17] van Halteren, H., Zavrel, J. and Daelemans, W. (2001) 'Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems', *Computational Linguistics*, pp. 199–229. doi: 10.1162/089120101750300508.
[18] *60 Incredible and Interesting Twitter Stats and Statistics* (no date). Available at: https://www.brandwatch.com/blog/twitter-stats-and-statistics/ (Accessed: 25 August 2020).
[19] langdetect (no date). Available at: https://pypi.org/project/langdetect/ (Accessed: 24 August 2020).
[20] googletrans (no date). Available at: https://pypi.org/project/googletrans/ (Accessed: 24 August 2020).
[21] Ramezani, M. and Fatemizadeh, E. (2010) 'Comparison of Supervised Classification Methods with Various Data Preprocessing Procedures for Activation Detection in fMRI Data', Computational Neuroscience, pp. 75–83. doi: 10.1007/978-0-387-88630-5_5.
[22] re — Regular expression operations — Python 3.8.5 documentation (no date). Available at: https://docs.python.org/3/library/re.html (Accessed: 24 August 2020).
[23] Python - Word Tokenization (no date). Available at: https://www.tutorialspoint.com/python_data_science/python_word_tokenization.htm#:~:text=Word%20tokenization%20is%20the%20process,for%20a%20particular%20sentiment%20etc. (Accessed: 25 August 2020).
[24] Code Faster with Line-of-Code Completions, Cloudless Processing (no date). Available at: https://www.kite.com (Accessed: 25 August 2020).
[25] Stemming and lemmatization (no date). Available at: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html (Accessed: 25 August 2020).
[26] Rahmatulloh, A. et al. (2019) 'Comparison between the Stemmer Porter Effect and Nazief-Adriani on the Performance of Winnowing Algorithms for Measuring Plagiarism', International Journal on Advanced Science, Engineering and Information Technology, p. 1124. doi: 10.18517/ijaseit.9.4.8844.
[27] 2. Accessing Text Corpora and Lexical Resources (no date). Available at: https://www.nltk.org/book/ch02.html (Accessed: 25 August 2020).
[28] Python Examples of nltk.sentiment.vader.SentimentIntensityAnalyzer (no date). Available at: https://www.programcreek.com/python/example/100005/nltk.sentiment.vader.SentimentIntensityAnalyzer (Accessed: 25 August 2020).

[29] API Reference — TextBlob 0.16.0 documentation (no date). Available at: https://textblob.readthedocs.io/en/dev/api_reference.html (Accessed: 25 August 2020).