Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach

Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach

Cristian R. Machuca[1*], Cristian Gallardo[1], Renato M. Toasa[2]

[1]Tomsk Polytechnic University, 30 Lenin Avenue, Tomsk, Russia
[2]Universidad Tecnológica Israel, Quito, Ecuador

*Email: cristian.machuca.mendoza@gmail.com

Abstract. In machine learning, a fundamental challenge is the analysis of data to identify feelings using algorithms that allow us to determine the positive or negative emotions that people have regarding a topic. Social networks and microblogging are a valuable source of information, being mostly used to express personal points of view and thoughts. Based on this knowledge we propose a sentiment analysis of English tweets during the pandemic COVID-19 in 2020. The tweets were classified as positive or negative by applying the Logistic Regression algorithm, using this method we got a classification accuracy of 78.5%.

1. Introduction

Among the most common viral infections that affect humans are the respiratory infections, which are caused by Human Respiratory Viruses (RVs) [1]. The best-known type of respiratory viral infection is the influenza or "flu", and every year causes between 250,000 and 500,000 deaths worldwide, being the H1N1 virus the most well-known variant [2]. One of the family of viruses that causes respiratory diseases are the coronavirus, which in humans infects the epithelial cells of the respiratory tract, being sometimes unnoticeable, but in some cases deadly, and can even affect other mammals and birds. There are several types of coronaviruses, the best-known are: The Middle East Respiratory Syndrome (MERS), the Severe Acute Respiratory Syndrome (SARS) and nowadays the Coronavirus Disease (COVID-19) [3,4].

The first cases of people having symptoms of infection in the respiratory tract caused by coronavirus occurred in mid-December, 2019. On December 31, 2019, the Wuhan Health Commission published information on cases about atypical pneumonia affecting patients coming from a local market in the city of Wuhan - China [5]. By late February, 2020, more than 4500 cases and more than 60 deaths related to COVID-19 had been confirmed outside of China [6]. On March 11, 2020, approximately 118,000 people were infected in 114 countries and 4,291 deaths had been confirmed, due to these alarming levels of severity and spread of coronavirus the World Health Organization (WHO) declared the COVID-19 disease as a pandemic [7-9].

The first COVID-19 impact analysis on humans revealed the severity of the infection. Among 67 patients, 3 (4.5%) were mild, 35 (52.2%) were moderate, 22 (32.8%) were severe and 7 (10.4%) were critically ill. The technique used to determine the severity level of the disease was the computerized tomography (CT) [10].

Artificial Intelligence (AI) is associated with the study of designing human or animal behaviour in artificial entities, specifically in learning and problem-solving issues [11]. In medicine, the use of artificial intelligence has been grouped into two branches: virtual and physical, with a wide range of systems such as information management, guidance of doctors in their treatment decisions and robots to help patients [12].

The use of machine learning techniques for the search of feelings contained in a text expressed in microblogging social networks is a fundamental point to understand people's perception of the impact that the COVID-19 pandemic has had at social, economic, political and technological levels. Based on the monthly analysis of English tweets since the beginning of the pandemic, the following article is presented.

The paper has been structured as follows. Section 2 presents an overview of the problems caused by COVID-19. Section 3 shows the methodology used to carry out the research presented. The experimental results are shown in section 4 and finally in section 5 the conclusions of this work are presented.

2. Problem Definition and Related Areas

The diseases that currently affect the world, especially which are classified as pandemic, cause serious problems to the population at all levels: economic, emotional, status, planning, politics, etc., in addition to the complexity of traditions, ethics, individual psychology and social behaviour of people. Therefore, it is required and necessary a people's attitudes analysis when adverse situations arise. The global crisis caused by the COVID-19 [13], has changed the global perception and way to deal with a large-scale disaster and has placed a significant psychological burden on people. Identifying people's reaction to this threat can provide important information on how society behaves and reacts to unwanted and unexpected situations, which can be positive or negative, currently the Internet and social networks have become powerful tools to access people’s opinions and comments on various topics [14].

With a large amount of information that can be found on the Internet and sometimes can be unreliable, it is necessary to use certain techniques to collect and analyse the existing data. Web scraping [15] is a technique for extracting structured information from a website in order to manage it as data for further analysis and visualization, through the use of bots.

Determining human behaviour has been one of the most relevant research topics of the modern era, and nowadays it is still a challenge that seems to have no effective solution, mainly because the human behaviour involves different areas of study such as philosophy, personality, social and environmental elements [16]. On the other hand, with the impact of the COVID-19 disease, new factors that affect the people's behaviour have been identified, such as poverty and economic dislocation, which have directly affected the reduction of compliance with safety protocols [17], other factors that are associated to the behaviour and the effectiveness of the crisis management are education level, income level, culture of the region and age [18].

According to the latest WHO reports [19], the number and severity of mental health problems will increase because of the long-term economic and social costs caused by the disease. Despite the risk, and due to the magnitude of the crisis, mental health needs are not receiving the required attention, something that has been exacerbated by the lack of investment and prevention in this area before the arrival of the pandemic. For this reason, analysing the public opinion can provide valuable information for understanding the evolution of people's emotions facing the pandemic.

2.1. Social Psychology and Microblogging Type Social Networks

In recent years, there is an increase in the use of social psychology in conjunction with various branches of computer science to analyse, understand and predict the behaviour of populations, through the people’s feelings and thoughts [20].

The social phenomena are influenced and caused by various factors such as motivations, emotions, attitudes, physical health, behaviours of others, moral perception, sense of justice and the radius of influence on people; which during many centuries was tiny in relation to today's globalized civilization. Nowadays, the power of influence of our closest relationships is often overshadowed by hundreds or thousands of people outside our city and even our country [21].

To analyse the behaviour of the population, one of the most widely used tools is Twitter, which is a micro-blogging type social network that allows users to express their thoughts or opinions using short phrases [22].

2.2. Natural Language Processing (NLP)

The difference between text mining and natural language processing must be taken into account. Text mining focuses on the discovery and extraction of information of interest within an unstructured text, whereas natural language processing aims to perform an extraction of a more complete meaning indicator from a text, trying to find out who, when, where, how and why an action was performed, to achieve it, NLP applies complex algorithms to perform different types of analysis such as morphological and lexical, syntactic, semantic, discourse integration and pragmatic [23].

2.3. Tokenizing

The process of tokenization is the division of a long text into sentences which will be delimited by punctuation marks showing the end, or by words creating a list that stores all the words in a text individually [24].

2.4. Stemming and Lemmatization

Stemming is the technique that identifies conjugated words and represents them in a unique way that expresses the same meaning and works with heuristics that seeks to cut out the words to standardize all conjugations and derivations. On the other hand, the Lemmatization technique applies a more complex analysis that through a word morphological analysis tries to find the base form of the conjugated words e.g. "am", "are", "is" is represented with its base form "be" [25].

2.5. Web Scraping

Web scraping refers to the use of applications to extract information from websites automatically or semi-automatically by simulating user navigation. Although the indexation of content by search engines performs similar operations, a notable difference is that web scraping performs a process of transforming the information of websites to achieve adequate storage in local devices for further analysis. Although there are many methods to carry out the information gathering process, nowadays for complex information recognition analysis purposes, a semantic analysis of contents is performed in order to obtain only the desired information [26].

3. Methodology

This section describes the knowledge areas involved in the analysis, as well as the technologies and processes used.

3.1. Tweet Collection, Cleaning and Analysis Process

This research is divided into two stages: the first consists of obtaining, cleaning and transforming text from Twitter and the second stage consists of developing a machine learning algorithm for the analysis and classification of tweets according to feelings.

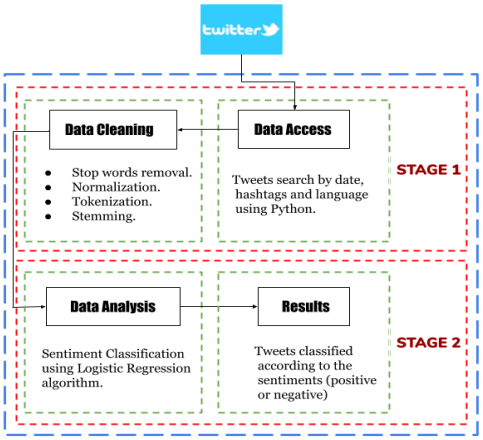

Figure 1. Flowchart of the twitter sentiment analysis classifier.

As shown in Figure 1 the first stage starts with Data Access; in this process the tweets were gathered using the Python programming language. The 50000 top English tweets during each month with the hashtag #coronavirus were downloaded. The data cleaning process started transforming each tweet to lowercase and removing all the punctuation marks, then the tweets were tokenized to facilitate the removal of non-English words and stop words, finally the words were stemmed and re-joined.

Once the tweets were clean, the second stage started, which implied the development of a machine learning algorithm to classify tweets according to feelings (positive or negative). In this case, the logistic regression algorithm was used, but before modelling the algorithm it was necessary to vectorize the tweets.

Machine learning algorithms only support numbers as input, thus the words must be coded as integers or floating-point values, this process is called feature extraction or vectorization.

The technique used to vectorize text documents for machine learning was TF-IDF (Term Frequency - Inverse Document Frequency), this is a technique to quantify a word within a text, thus weights will be assigned to each one of the words, which means the importance of the word in the document, and is calculated as follows:

*TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)*

Where TF is the frequency of a word in a document, while IDF is the inverse of the number of documents where the word is present.

Each document and word has its own TF, and is given by:

TF = number of times a word appears in a document / total number of words in the document

$$TF = t/n$$

The IDF measures the informativeness of a word, or how rare is to find a word in a certain number of documents:

IDF = total number of documents / number of documents where a word is present

In case a word is not contained in any document, DF will be equal to 0, as dividing by 0 is undefined, the last equation was modified, having:

IDF = log (total number of documents / (number of documents where a word is present + 1))

$$IDF = log(N/(DF+1))$$

Finally, combining the equations, it is possible to get the TF-IDF score:

$$TF\text{-}IDF = (t/n) * log(N/(DF + 1))$$

Once the vectorization process finished, we proceeded to the tweets classification by sentiments (positive or negative), for this purpose we previously trained the model, for the algorithm modelling we used a dataset that contains approximately 1600000 tweets, which were classified as positive or negative, we selected the binary logistic regression algorithm, because the dependent variable only has two possible outputs, positive or negative. After training the model, the confusion matrix seen in Figure 2 was obtained.

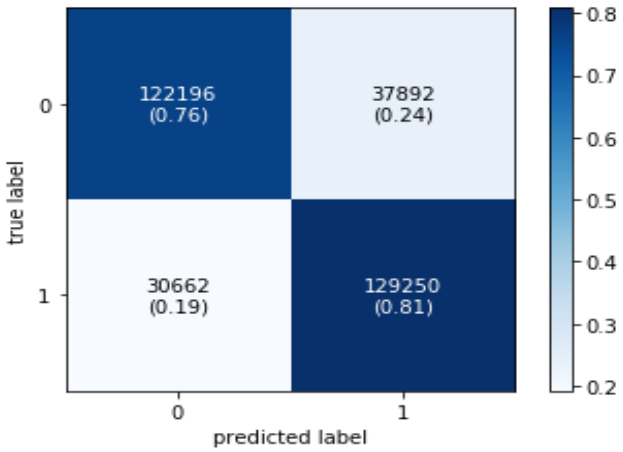

Figure 2. Confusion matrix for Binary Logistic Regression model (overall accuracy 78.57%).

4. Experimental Results

As it was described earlier, each month 50,000 top tweets were downloaded. The analysis was carried out from January to July, 2020, the search criteria was the hashtag #coronavirus.

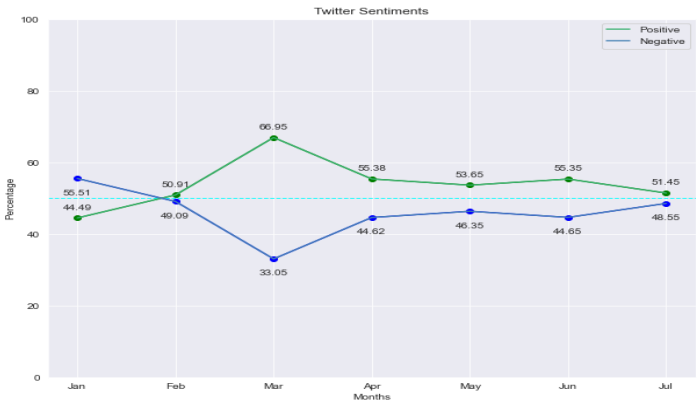

Figure 3. Percentage of positive and negative tweets month by month.

As shown in Figure 3, only during the month of January tweets were mostly negative, being the month of March the most positive, coinciding with the adoption of quarantine as a preventive measure by most European and Asian countries.

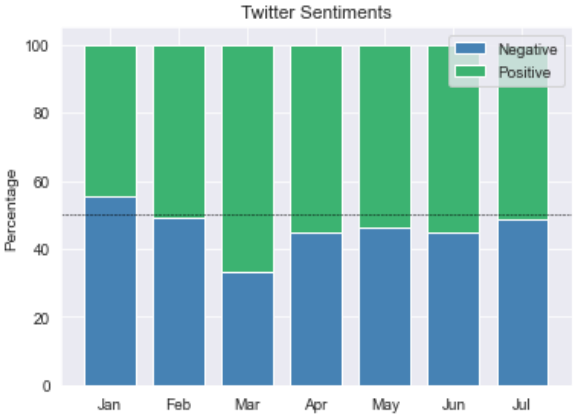

Figure 4. Stacked chart of the tweets sentiment analysis.

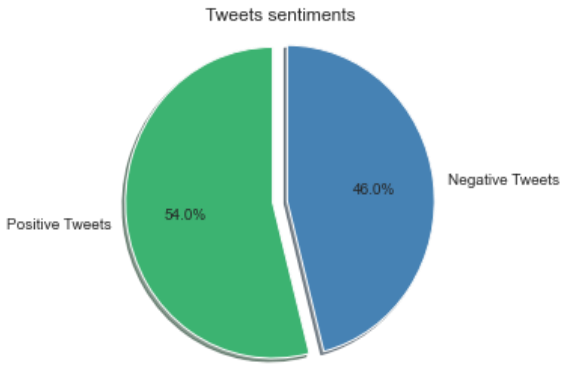

Figure 5. Pie chart of the overall tweets sentiment analysis.

There is a great similarity in the amount of negative and positive tweets in terms of percentage (See Figure 4), being the months of February and July when positive and negative feelings present similar percentages.

According to Figure 5, during the first seven months of the year, the number of positive tweets was slightly higher, however, there is a trend from April to July to equalise the number of positive and negative tweets, which indicates that approximately half of the people who express their opinions on Twitter have negative feelings and are pessimistic about the pandemic.

5. Conclusions

Taking into account that the COVID-19 disease is global health problem and has affected most countries and their economies, this paper focuses on analysing people's reaction to the pandemic. The main goal of the research is to deduce whether the sentiment of the public opinion is positive or negative by applying machine learning algorithms and NLP techniques. Despite the fact that the analysis found variation of opinions, it seems that people mostly remain positive about the pandemic, January is the only month in which negative thoughts predominated, March is the month when the COVID-19 disease was declared as a pandemic and many countries started to apply care measures and safety protocols, which coincides with the rise of positive thoughts. To summarize, 54% of the users showed positive feelings and 46% of the users showed negative feelings.

The scheme proposed within the methodology has a universal approach and changing the data source allows to analyse and obtain the frequency of the desired results within the examined text, guaranteeing the replication of this methodology in different works with similar characteristics.

References

[1] Schaffer K, La Rosa A M, and Whimbey E 2010 Chapter 162 - Respiratory viruses.
[2] De Gascun C F, Carr M J, and Hall W W 2010 Chapter 161 - Influenza viruses.
[3] King A M Q and Adams M J and Carstens E B and Lefkowitz E J 2012 *Family - Coronaviridae*. Elsevier.
[4] Korsman S N J, van Zyl G U, Nutt L, Andersson M I, and Preiser W 2012 *Human Coronaviruses* (Churchill Livingstone)
[5] Ralph R, Lew J, Zeng T, Francis M, Xue B, Roux M, Ostadgavahi A, Rubino S, Dawe N, Al-Ahdal M, Kelvin D, Richardson C, Kindrachuk J, Falzarano D, and Kelvin A 2020 2019-nCoV (Wuhan virus), a novel Coronavirus: human-to-human transmission, travel-related cases, and vaccine readiness *The Journal of Infection in Developing Countries*, **14**(01), 3-17
[6] World Health Organization (WHO). 2020 *Coronavirus disease 2019 (COVID-2019) Situation Report-39*. Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200228-sitrep-39-covid-19.pdf
[7] World Health Organization (WHO) 2020 *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19* - 11 March 2020. Retrieved from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
[8] Ducharme J 2020 *World Health Organization Declares COVID-19 a "Pandemic" Here's What That Means 2020*. Retrieved from https://time.com/5791661/who-coronavirus-pandemic-declaration/
[9] BBC News 2020 Coronavirus confirmed as pandemic by World Health Organization. March 11, 2020 Retrieved from https://www.bbc.com/news/world-51839944
[10] Zhong Qi, L. 2020 CT imaging features of patients with different clinical types of COVID-19 *Journal of Zhejiang University (Medical Science)*, **49**(2), 198
[11] Luger G F 2008 *Artificial Intelligence: Structures and Strategies for Complex Problem Solving.* (Addison-Wesley)
[12] Hamet P, and Tremblay J 2017 Artificial intelligence in medicine. *Metabolism: Clinical and Experimental*, 69S, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011
[13] Fauci A S, Lane H C, and Redfield R R 2020 Covid-19 - Navigating the uncharted, *New England Journal of Medicine* **382**(13) 1268–1269
[14] Holford M E, Cheung K-H, Zheng M, Krauthammer M, King T, and Chute C G, *Advances in Semantic Web* p 248
[15] Glez-Pen‹ D, Lourenco L, Lopez-Fernandez H, Reboiro-Jato M, and Fdez-Riverola F *Web Scraping Technologies in an API World*
[16] Arru M, Negre E 2017 People behaviors in crisis situations: Three modeling propositions. *14th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2017)* (May 2017, Albi, France) pp.139-149. ffhal-01729057f

[17] Wright A L, Sonin K, Driscoll J, and Wilson J 2020 Poverty and Economic Dislocation Reduce Compliance with COVID-19 Shelter-in-Place Protocols *Journal of Economic Behavior & Organization* 180, 544-554
[18] Bezerra A C V, da Silva C E M, Soares F R G, and da Silva J A M 2020 Fatores associados ao comportamento da população durante o isolamento social na pandemia de COVID-19 *Cien Saude Colet*, 25(1) 2411–2421
[19] World Health Organization (WHO) 2020 Facing mental health fallout from the coronavirus pandemic May 29, 2020 Retrieved from https://wwwwhoint/news-room/feature-stories/detail/facing-mental-health-fallout-from-the-coronavirus-pandemic
[20] Baron R, and Byrne D 1987 *Social psychology: Understanding Human Interaction*, 5th ed (Allyn & Bacon)
[21] Fiske S T, Gilbert D T, and Lindzey G 2010 *Handbook of Social Psychology Hoboken* (NJ, USA)
[22] Juan G et al 2018 Harvesting Opinions in Twitter for Sentiment Analysis *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, **2018**, pp 1–7
[23] Kao A and Poteet S R 2007 *Natural Language Processing and Text Mining* (Springer, London)
[24] Hassler M and Fliedl G 2006 Text preparation through extended tokenization WIT *Transactions on Information and Communication Technologies* 37, 13-21
[25] Gupta R, Jivani A G 2017 Analyzing the stemming paradigm En *International Conference on Information and Communication Technology for Intelligent Systems* (Springer, Cham) pp 333-342
[26] Glez-Peña D, Lourenco A, López-Fernández H, Reboiro-Jato M, and Fdez-Riverola F 2013 Web scraping technologies in an API world *Briefings in Bioinformatics*, **15**(5), 788–797