Twitter Text Data from #Covid-19: Analysis of Changes in Time Using Exploratory Sentiment Analysis

Eryka Probierz1*, Adam Gałuszka1 and Tomasz Dzida2

1 Department of Automation and Robotics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
2 Warsaw Stock Exchange, Książęca 4, 00-498 Warsaw, Poland

*Email: erykaprobierz@gmail.com

Abstract. The aim of this study is to analyse the characteristics of polish Covid-19-focused twitter posts identified via Covid-19 hashtags. The use of the popular service gives us an opportunity to monitor changing trends in users' opinions, and changing focus of the most popular topics in the Covid-19-oriented tweets. As part of the analysis, an exploratory sentimental analysis was used to identify the main topics, emotional characteristics, and possible changes over time in the tweeter statements. The emotional characterization of the statements allowed to classify them as negative, neutral or positive. The analysis used wide range of classical text mining, sentiment analysis and latent semantic analysis methods. The use of the methods allowed obtaining the most accurate picture of the tweets. The obtained results indicate the importance of social media in creating opinions.

1. Introduction

The coronavirus pandemic, which broke out in 2020 and spread all over the world, is a topic that humanity is constantly fighting against and trying to find the best solution. The changes in societies that have been forced by adaptation to the new reality have disturbed a certain world order and order known so far [1-3]. People are confronted with the need to reorganize their work, private or family life. Every day they face difficult choices, the consequences of which may have long-term consequences. They take actions or abandon them, not only with their own plans and goals in mind, but also with the need to adapt to the existing reality. At the time of writing this article, no solution with a broad spectrum of effectiveness is yet known. The world, after the fight against the first wave of infections and the temporary unfreezing of society, starts to struggle with the second wave, the intensity and aggressiveness of which seems to be even more acute [4,5].

However, regardless of the changes in the functioning of societies, one thing has remained constant, namely the broad spectrum of using social media as a medium through which one can express their opinions, fears or concerns. In which one can say what one thinks, what one is angry with and what one is upset about. In which you can refer to the actions of other people or confront your opinions with those of hundreds or thousands of other users [6]. One of the available social media is Twitter, which gives the opportunity to publish short text messages with the hashtags used. A culture that promotes speed and time offers a tool to collect information in the shortest and most accessible form, in Tweet form [7,8].

Just as the coronavirus pandemic has become widespread, so is the use of social media. In an era of widespread isolation, avoidance of human contact and communication with the use of technology that
transmits data, the power of social media has increased and constitutes an important source of information about society and cultures [9]. About the difficulties they are struggling with, about the widely understood opinions, about moods. Social media have become a metaphor for a mirror that reflects what is perceived as extremely important in society at a given moment. It is important enough that the hashtag that a given Tweet is carrying refers to this phenomenon [10,11].

2. Data
In Poland, the number of Twitter users has reached 4 million and is constantly growing. At the time of writing this article, the daily number of morbidity and deaths related to the coronavirus is also growing. The aim of this article is to analyze a selected pool of Tweets, which were provided with appropriate hashtags associated with the coronavirus and to conduct a sentimental analysis on the obtained data set.

The obtained data will be assigned to a time category which will correspond to three stages of the pandemic, i.e. first wave of the disease, unfreezing and second wave of the disease. For each of these three subgroups keywords will be distinguished and the frequency of their occurrence will be examined. In addition, analyses will be carried out on the differences in the occurrence of given keywords, depending on the stage of the pandemic in which it was published. Sentiment analysis methods will be used to examine whether the published sentences are neutral, negative or positive and how this distribution is presented not only for the whole group of analyses but also for individual key words.

The study analyzed Tweets marked with the following hashtags in Polish:

- #corona
- #coronavirus
- #Coronavirus
- #covid
- #covid-19
- #Coronaviruspolska
- #covid-19PL
- #covidPl

121 412 sentences from the reviews were analyzed, which gave a total of 57 454 tweets. Next step was to define tweets aspects, which are general facets about topic of tweet. Basing on other works, a two-step approach was applied. First it was done a manual inspection of tweets and tweets hashtags to identify most common aspects. Then it built a system capable of predicting tweet aspects in an automated way. System was built with use of supervised learning techniques. 11 tweets aspects were included in analysis and it was:

- Vaccine (i.e. Coronavirus vaccine becoming more and more realistic.)
- fake news (i.e. The most popular fake news about coronavirus.)
- influenza (i.e. Coronavirus is a more severe version of the influenza.)
- ministry of health (i.e. The Minister of Health has a coronavirus.)
- quarantine (i.e. Quarantine for people arriving from abroad.)
- restrictions (i.e. New restrictions for people under 18 years old due to coronavirus.)
- masks (i.e. The difference between a protective mask and a visor.)
- mortality (i.e. Mortality from coronavirus of the elderly.)
- complications (i.e. Coronavirus and lungs – complications.)
- conspiracy (i.e. Coronavirus suppressing influenza as a conspiracy theory.)
- symptoms (i.e. Most common symptoms of coronavirus.)

To predict manually invested aspects of Tweets it was used model of logistic regression in Statistica 13.3. The model yielded a macro F1 0.491, and micro F1 0.694. After literature review, an approach including convolutional neural network was applied using RapidMiner. Best performance was achieved using CNN model and results are in Table 1.
Table 1. Accuracy levels of the tweets prediction models

|                      | Logistic regression | Convolutional Neural Network |
|----------------------|---------------------|-------------------------------|
| Accuracy             | 0.694               | 0.880                         |
| Macro F1 score       | 0.491               | 0.864                         |
| Micro F1 score       | 0.694               | 0.880                         |

3. Sentiment Analysis
Based on the reviewed solutions, it was decided to choose VADER - Sentiment Analysis as the preferred solution to analyze tweets aspects. Sentiment Analysis allows to measure emotional language characteristics using linguistic methods, NLP and text analysis. Reactions are analyzed and evaluated in a dictionary with an appropriate weight. Sentiment analysis can be performed in two approaches. The first approach assumes using only statistical methods for text analysis. It ignores the order of words or the context of the statement. The second approach integrates statistical and linguistic methods in order to better understand the analyzed statement. The obtained data can be divided into two categories, the first one assigns a neutral, positive or negative label. The second one adds a scale to the above categories, so the word can be more or less positive/negative.

4. Results
First analysis contains information about the number of negative, neutral and positive sentences in all data set of tweets. From 121412 tweets 67144 was negative, 36252 was neutral and only 18016 was positive. The second stage analyzed the occurrence of negative, neutral and positive sentences in the highlighted tweets aspects. The obtained results are presented in Table 2 and Figure 1.

Table 2. Number of sentences assigned to particular application aspects.

| Application | Negative | Neutral | Positive |
|-------------|----------|---------|----------|
| Vaccine     | 4910     | 3647    | 6457     |
| fake news   | 7483     | 2541    | 2561     |
| Influenza   | 8891     | 1254    | 674      |
| ministry of health | 4574 | 2074 | 1287 |
| quarantine  | 8744     | 3645    | 699      |
| Restrictions| 6988     | 2521    | 783      |
| masks       | 2345     | 4514    | 2415     |
| mortality   | 5674     | 6489    | 489      |
| complications| 6692 | 3647 | 1087 |
| conspiracy  | 4156     | 1842    | 652      |
| Symptoms    | 6687     | 4078    | 912      |
To check whether percentage differences between tweet aspects in different emotional rates are statistically significant, the difference between the two structure indicators was applied. Results are in Table 3.

Table 3. Differences between tweets aspects.

| Aspect     | Negative | Neutral | Positive | Neg. vs neu. | Neu. vs pos. | Neg. vs pos. |
|------------|----------|---------|----------|--------------|--------------|--------------|
| vaccine    | 7%       | 10%     | 36%      | p<0.001      | p<0.001      | p<0.001      |
| fake news  | 11%      | 7%      | 14%      | p<0.001      | p<0.001      | p<0.001      |
| influenza  | 13%      | 3%      | 4%       | p<0.001      | p<0.01       | p<0.001      |
| ministry of health | 7% | 6% | 7% | p<0.01 | p<0.01 | ----------- |
| quarantine | 13%      | 10%     | 4%       | p<0.001      | p<0.001      | p<0.001      |
| restrictions | 10%  | 7%      | 4%       | p<0.001      | p<0.001      | p<0.001      |
| masks      | 3%       | 12%     | 13%      | p<0.001      | p<0.01       | p<0.001      |
| mortality  | 8%       | 18%     | 3%       | p<0.001      | p<0.001      | p<0.001      |
| complications | 10%  | 10%     | 6%       | ----------- | p<0.001      | p<0.001      |
| conspiracy | 6%       | 5%      | 4%       | p<0.001      | p<0.01       | p<0.001      |
| symptoms   | 10%      | 11%     | 5%       | p<0.01       | p<0.001      | p<0.001      |
| total      | 67144    | 36252   | 18016    |              |              |              |

5. Conclusion and Limitation of the Study
The obtained results indicate the importance of analysing individual tweet aspects in order to understand the complexity of the problem of coronavirus tweets in Poland. The results obtained indicate that with the exception of two pairs, all differences in the percentage share were statistically significant. This means that depending on whether a given sentence was assigned to a neutral, negative
or positive category, the percentage share of the individual aspects distinguished was so different. It is pointed out, however, that the results obtained should be generated with great caution, due to the constantly increasing collection of tweets related to the coronavirus in Poland. The above analysis indicates, however, that it is possible to analyze the emotional character of the entries based on data from social media.

Acknowledgments
The work of Eryka Probierz was supported in part by the European Union through the European Social Fund as a scholarship under Grant POWR.03.02.00-00-1029, and in part by the Silesian University of Technology (SUT) through a grant: the subsidy for maintaining and developing the research potential in 2020 for young researchers in data collection and analysis under Grant 02/060/BKM20/0012. The work of Adam Galuszka was supported by the SUT under Grant 02/060/BK_20/0007 (BK-276/RAU3/2020) the subsidy for maintaining and developing the research potential, in 2020. This work was supported by the (GeCONiI) Upper Silesian Centre for Computational Science and Engineering through The National Centre for Research and Development (NCBiR) under Grant POIG.02.03.01-24-099/13.

References
[1] Chen E, Lerman K, and Ferrara E 2020 Covid-19: The first public coronavirus twitter dataset arXiv preprint arXiv:200307372
[2] Chen E, Lerman K, and Ferrara E 2020 Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set JMIR Public Health and Surveillance, 6(2), e19273
[3] Sharma K, Seo S, Meng C, Rambhatla S, Dua A, and Liu Y 2020 Coronavirus on social media: Analyzing misinformation in Twitter conversations arXiv preprint arXiv:200312309
[4] Budhwani, H, and Sun, R 2020 Creating COVID-19 Stigma by Referencing the Novel Coronavirus as the “Chinese virus” on Twitter: Quantitative Analysis of Social Media Data Journal of Medical Internet Research, 22(5), e19301
[5] Saire J E C, and Cruz J F O 2020 Study of Coronavirus Impact on Parisian Population from April to June using Twitter and Text Mining Approach medRxiv
[6] Aguilar-Gallegos N, Romero-García L E, Martínez-González E G, García-Sánchez E I, and Aguilar-Avila J 2020 Dataset on dynamics of Coronavirus on Twitter Data in Brief, 105684
[7] Saleh S N, Lehmann C U, McDonald S A, Basit M A, and Medford R J 2020 Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter Infection Control and Hospital Epidemiology 2020 1-8
[8] Rajput N K, Grover B A, and Rathi V K 2020 Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic arXiv preprint arXiv:200403925
[9] Prabhkarakaila D, and Prasad D A 2020 Informational flow on Twitter—Corona virus outbreak–topic modelling approach International Journal of Advanced Research in Engineering and Technology (IJARET), 11(3)
[10] Pastor C K 2020 Sentiment Analysis of Filipinos and Effects of Extreme Community Quarantine Due to Coronavirus (Covid-19) Pandemic Available at SSRN 3574385
[11] Jelodar H, Wang Y, Orji R, and Huang H 2020 Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach arXiv preprint arXiv:200411695