STUDIA UBB SOCIOLOGIA, 65 (LXV), 1, 2020, pp. 111-130
DOI: 10.2478/subbs-2020-0005

THE REBELLIOUS SOCIAL NETWORK REACTION TO COVID-19

ŞTEFANA CIOBAN[1], DRAGOŞ VÎNTOIU[2]

ABSTRACT. Gathering social media content and analysing the heavy and unstructured text coming from posts, comments and reactions can come as a powerful tool in understanding how people react to the information they receive. In this article we present the results from a social media analysis of 10771 headlines, with their subsequent text bodies and comments posted in a subreddit destined for Romanians during the state of emergency declared in Romania, from March 16 to May 15, 2020. Our objective was to model the main topics debated by this targeted population of people that tend to use Reddit to discuss current issues and to identify the sentiment polarity towards these topics. As expected, Romanians are mostly concerned with their social condition in the context of the pandemic caused by CoVID-19, as our research has revealed a word frequency for the term "Coronavirus" prominently higher than any other preferred term. However, the analysis brings up a surprising turnaround as the overall sentiment of the text posted in this dataset is predominantly neutral with a higher frequency of positive posts compared to the negative ones. This was unforeseen by our initial expectations: a natural tendency to more negative posts than positive considering the context of the chosen study period. Moreover, when compared to the time series of the CoVID-19 infections and caused deaths in Romania, spikes of extremely high or low mean sentiment scores per day can be correlated to the fluctuations of the declared cases. Not only does this bring us closer to understanding the social impact of CoVID-19 in the current context, but the outcome of this analysis can be easily extrapolated for further investigations upon other social networking tools or for more in-depth analysis on our studied corpus.

Keywords: social media analysis, sentiment polarity, topic modelling, CoVID-19, state of emergency

[1] *Masters in Complex Data Analysis, Faculty of Sociology and Social Work, Babeș-Bolyai University Cluj-Napoca, e-mail: stefanacioban@yahoo.com.*

[2] *Masters in Complex Data Analysis, Faculty of Sociology and Social Work, Babeș-Bolyai University Cluj-Napoca, e-mail vintoiu.dragos@gmail.com*.

Introduction

Public opinion is a key point in analysing how the information is spread and impacts people's lives. We refer to the public opinion as it is defined by Krippendorff (2005), in which opinion implies an independent, unpredictable and cognitively autonomous exercise of one's mind, while public, a noun coming from Latin and meaning "people", refers to a collective and in our case, a collective which exercises their ability of having opinions in the public sphere.

With the growing popularity of social networking, the heavy content of posts, reactions and comments allows us to analyse what people are most concerned of and their reaction to such concerns (Fersini, 2017). This, in turn, brings us closer to understanding how is it that the society expresses concern about an issue, in our case, the perception over a period of time about CoVID-19 social, medical, economic, and political implications. The way we studied public opinion via social networking is an expression of favouring positive feelings or indulging negative feelings in people's lives.

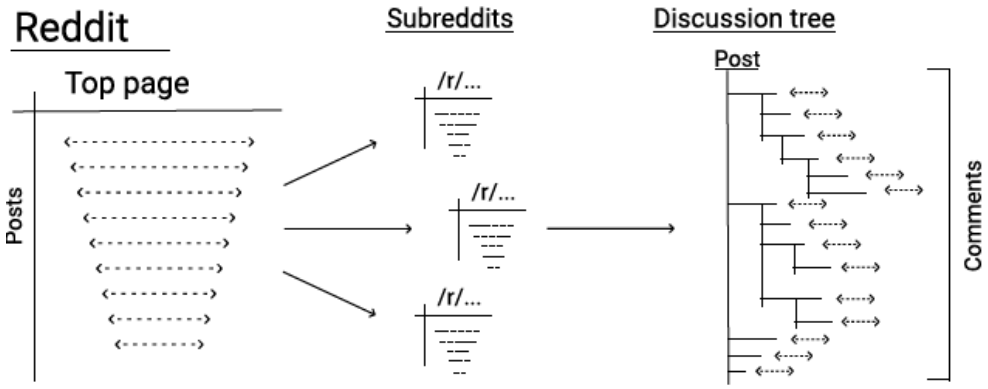
In this study, we have analysed almost 7.5 million words from Reddit posts to determine the polarity and the inclination towards a popular subject. CoVID-19 was the aimed subject taken into consideration due to the economic, political, and social implications that have boomed during the last months (Swasey, Winter, & Sheyman, 2020). We have not, however, targeted only the posts who were related to this subject, but rather modelled the topic out of the text corpus to see whether this situation was discussed via social media. On the 16th March 2020, the president of Romania announced that the country is under a 'state of emergency', implying a series of restrictions and new rules that threatened to change the lives of many citizens. On the 15th May 2020, the country switched from state of emergency to 'state of alert' with fewer restrictions.

In this context, our study focuses exclusively on the Reddit (Reddit, 2019) text posted in a Romanian subreddit during this period and aims to the understanding of how these changes and the way they were communicated have affected Romania's young to middle-aged population. Reddit is a social networking tool intensively used by young adults and an increasing audience consisting of young and middle-aged people. It is a platform that combines web content, forums, and social news, organized in subreddits. Submissions get upvoted by other users, which makes it easy to identify the popularity of the content posted by a user (Mediakix, 2018). Discussion trees (Figure 1), as described by Alexey N. Medvedev, Renaud Lambiotte and Delvenne Jean-Charles (Medvedev, Lambiotte, & Delvenne, 2018) are a form of organizing the posts and comments people submit and contours the life of a community.

This aspect further implies important public opinion factors: liking, supporting, having attitude, and expressing one's beliefs upon something, which are debated on this social media platform in a dynamic, forum-like exchange of opinions. Reddit is by itself a social intrument of acting in one's beliefs since the posts get upvoted and engage people into giving feedback, hence they form an act of drawing attention (Medvedev, Lambiotte, & Delvenne, 2018). This empowered us in choosing Reddit as a platform of public opinion amongst so many others available online.

Moreover, Reddit users are pseudonymous, and there is more likely for them to express without constraints because they use Reddit services without displaying their real name. What might come as a surprise is that the platform collects personal data such as: email and IP address, private conversations, reactions, and posts. However, the information targeted in this article is the text corpus coming from the posts and their comments, and no user-related data has been analysed (Reddit, 2020). Once again, we have chosen Reddit for our study because it is one of the largest web services, it has a dominant database with more than 300 million users, and in the last 6 months they gathered over 1.5 billion visits (SimilarWeb, 2020). The platform comes with an API destined for the fast and easy collection of titles, bodies and comments for defined periods of time (Reddit, 2018).

Figure 1. Schema of the structure of the Reddit platform: top page organized in subreddits and discussion trees



Source: Medvedev, Lambiotte, & Delvenne (2018).

This article first covers an overview of the existing literature discussing the matter of how people are generally reacting during a social crisis, which are the feelings they tend to share and how they are adapting to an unusual situation that is threatening to alter most spheres of the society. A particular focus is shaped around the times of pandemics, and around what happens when people tend to seek information and debate around the disease causing the pandemics in the first place. The main research questions are stated in this context followed by a section which targets the methodology of the analysis that drives to answering them. The methodology touches all stages of the analysis, illustrating how the results were driven and which were the limitations of the open source technologies we have used. This in turn is followed by an interpretation of the results gathered at all stages of the analysis: data preprocessing, word frequency analysis, sentiment evolution in time and topic modelling. These interpretations are wrapped together in the discussion section where our findings are theorized along the sociological findings together with the practical application of the text mining techniques applied on large sets of unstructured datasets coming from social media debates. In the last section, the conclusions of the study are drawn. Here we also harvest some of the future research directions instigated by our study.

Shared feelings during a social crisis

During a pandemic, it is normally expected that negative feelings such as uncertainty, fear, anxiety, or worry would arise among societies (Dingwall, Hoffman, & Staniland, 2013). In this way stress levels increase in individuals and expose more sensitive ones to more serious disorders (Shigemura, Ursano, Morganstein, Kurosawa, & Benedek, 2020). This could be explained by the fact that generally people have easy access to all shapes of information, including inaccurate and exaggerated news that circulates over the Internet and that intentionally or not increases the sense of fear and uncertainty (Ornell, Schuch, Sordi, & Kessler, 2020). Nonetheless, media plays an important role, as by spreading the news about the disease and its potential implications, it causes a sense of danger among the public, more than simply increasing its awareness. In this way, people normally perceive the problem in an alienated manner which differs from their own experience or knowledge. A serious effect can fall on "the sense of 'trust' that the society has in the face of mass problems", meaning that in such conditions people are starting to doubt the governing organs of the country as well as the states' official sources of information (Kasapoglu & Akbal,

2020). This in turn results in a weakening sense of security and hope for the future in the population of the country, ending up in a general negative feeling.

This causes the pandemics to turn from a natural problem to a social one as its effect is sensed in most of the spheres of the society: family, education, workplace and employment, public places, and open-air activities (Kasapoglu & Akbal, 2020). What causes this uncommon situation to be sensed as an instigator for fear and uncertainty is its effects on most social aspects in people's lives. Once the social distancing becomes mandatory, all activities that were done in groups have to be stopped (De Vos, 2020). This could either involve leisure activities or work or school related ones, all of which are part of our day to day lives. Once this significant part of our lives has to be readapted to the new reality we are facing, this triggers an entire chain of causalities to arise and to instigate uncertainty in the population. Not only does the disease itself cause an epidemiological impact, but also the fear of the implications of the disease is easily spread causing an impact in the social relations and consequently in their public discussions and debates.

Social media can be a powerful data source offering us a wide range of possibilities to discover previously unknown information about what people are discussing (Usai, Pironti, Mital, & Mejri, 2018). They can either be using it as means of expressing their opinion, or as a reaction to other people's opinion or simply to communicate with the others. Shaping this information using novel text mining techniques can help us to better understand the impact that a major event can have on the society (Ampofo, Collister, O'Loughlin, & Chadwick, 2015). Such major event is, in our case, the pandemic caused by CoVID-2019 and its implications on most sociological layers.

We should, however, be aware that this novel type of data comes in an unstructured format and in heavy quantities which require careful transformations to allow a fair interpretation of the information and the consequent creation of knowledge. The problem we are facing is the lack of consistent research in mining social data, therefore the full potential of the advances of text mining tools towards sociological studies has not yet been reached (Ristoski & Paulheim, 2016; Usai, Pironti, Mital, & Mejri, 2018; Gaspar, Pedro, Panagiotopoulos, & Seibt, 2016). What is more, the process of data construction itself yields a transparent delivery in a well-defined research framework (Mutzel, 2015).

Recent studies on social media data understanding how society reacts under the stress of a social crisis are given by Gaspar, Pedro, Panagiotopoulos, & Seibt (2016) and Dawn Breslin, Enggaard, Blok, Gårdhus, & Pedersen (2020). The latter employs an analysis of the same situation as the one studied in this article: the reaction to the pandemic and consequent lockdown caused by CoVID-2019. Both studies reveal that the studied countries have gone through

real stressful times which has provoked a combination of negative feelings in their population. On the other hand, Zhong et al. (2020) are arguing that, in China, the crisis deployed by CoVID-2019 is perceived in an optimistic, hence positive way by the people who "have appropriate practices towards COVID-19" or simply, who are more knowledgeable about the virus.

In this context, our main concern is the following: how did the Romanians who are expected to be well informed react to the pandemic caused by CoVID-2019? We took the study case of the Reddit social media reaction during the lockdown as the lead dataset to be studied. Our expectations were to identify a special focus on the discussion around this virus and its implications in the social life of the individuals, shared among the social network platform studied. Around this topic, we have also expected a tendency towards panic or fear, hence proof of negative feelings towards the causalities of the pandemic. Using the results of this analysis, answers to the following questions were sought:

a) Is social media a means of expressing powerful feelings towards the context of a social crisis?

b) Is the general sentiment identified in the text from the social media during the most restrictive times during the pandemic revealing fear or more of a positive feeling, such as optimism?

Methodology

We have followed the widely used data analysis steps for Natural Language Processing for our research dataset, using Python (Python, 2014) as a data science tool for fast analytics and Tableau (Tableau, 2009) to display the results (Deepa, Manjunath, & Ravindra, 2019), (Subramanian, 2019). First, we have preprocessed our dataset in terms of text cleaning, translations, and lemmatization, and second, we have counted the frequencies of all words' occurrences from the posts and comments analysed. We have taken both Romanian and English translations into account. Last, we have computed the sentiment polarity score and compared the results with the official CoVID-19 infections and deaths recorded per day (Portal, 2020). All steps of our methodology are represented in a series of Jupyter Notebook files published in our GitHub repository (Cioban & Vîntoiu, 2020).

Data containing the text of the posted titles, text bodies, and their comments, together with the number of comments and scores for each post were extracted using Reddit's available endpoint for data collection (Reddit, 2018). In Table 1, the descriptive statistics of the explored dataset are shown. The main variables of the study are the Reddit-computed score of each posted headline, the number of subsequent comments for each post, the computed length of words for the posts and comments in both English and Romanian and

the computed Vader (Hutto & Gilbert, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, 2014) score for each post translated in English. As seen from the descriptive statistics table, more text is concentrated in the comments than in the posts. This comes as expected as posts are generating a high number of comments, in this case leading to a maximum of 1098 comments for one single post. The top rated scores as per the Reddit scoring system (Reddit, 2019) seem to be getting close to 3000, whilst the least popular ones even get a score of zero points. As the mean of the number of words used in the comments is very high compared to the word length of posts, we have used the Romanian word frequencies from both the comments and corresponding posts to understand which are the terms preferred by Romanians over Reddit during the lockdown. In terms of the Vader polarity score, the mean seems to be close to zero, with a standard deviation of approximately 0.4 leading to spikes of positive and negative feelings identified among the posted headlines. We have extracted and analysed a total number of 10771 headlines and their subsequent 217815 comments available from the subreddit "Romania" (Reddit, 2019), during the state of emergency declared from March 16 to May 15, 2020. The analysed period contains 59 days, having computed a mean polarity score for each which was compared to the number of CoVID-19 infections and caused deaths (Table 2). We avoided using any specific tools to filter out the threads that were not discussing around the virus, to see if one of our main expectations was met: people discussing mostly around the implications of CoVID-19.

Table 1. Descriptive statistics of the titles, bodies, and comments in the Reddit dataset from the Romania subreddit retrieved between 16 March 2020 and 16 May 2020

	Reddit score	**Number of comments**	**Posts (Romanian) word length**	**Comments (Romanian) word length**	**Posts (English) word length**	**Posts (English) Vader polarity score**
mean	59.54	20.22	37.52	661.55	34.65	0.016
standard deviation	183.7	47.76	116.56	1830.01	129.09	0.4
minimum	0	0	0	0	0	-0.997
1st quartile	1	1	7	45	7	-0.077
2nd quartile	5	8	15	135	15	0
3rd quartile	40	23	27	586.5	27	0.226
maximum	2878	1098	2559	42389	7384	1

Source: personal computations based on the data from Reddit (2019).

Table 2. Descriptive statistics for the daily CoVID-19 registered cases and caused deaths between 16 March 2020 and 16 May 2020

	CoVID-19 cases	**CoVID-19 caused deaths**
count	59	59
mean	265.51	16.98
standard deviation	126.34	11.18
minimum	17	0
1st quartile	190	7
2nd quartile	278	19
3rd quartile	346.5	25
maximum	523	42

Source: personal computations based on the data from Portal (2020).

Considering the low availability of natural language processing support existent for the Romanian language, we have translated all bodies and titles' text to English using Yandex API (Yandex Technologies, 2019). This endpoint allows translations of up to 5 million words per day meaning that we were restricted to only translate the body and titles corpus, as the high volume of text in the comments was exceeding the allowed limit for text translation. The critical preprocessing required by any translation toolkit was to clean all text from special characters, emoticons, and URLs. Due to this requirement, an important aspect of the dataset was omitted: the reaction expressed via emoticons which is nowadays widely used in social media. Still, considering the forum-like expression of text typical to Reddit, text is usually preferred by users instead of emoticons as a means of transmitting reactions to other people's posts or comments (Medvedev, Lambiotte, & Delvenne, 2018).

All data was saved to a text corpus and analysed in terms of word frequency. We used the Bag of Words technique (Zhang, Jin, & Zhou, 2010). Therefore, we have constructed a dictionary consisting of the top 5000 words and used it for the word vectorization step to aid in visualizing the results and model the two main topics from the studied Reddit data. To model the word embeddings, we used Python's library scikit-learn, CountVectorizer (Borovikova, 2011).

To investigate the polarity of the posts, we have used the Vader rule-based lexicon (Hutto & Gilbert, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, 2014) to compute a compound sentiment score for each post, generically defined here as the titles joined with their corresponding body text. Vader was chosen for its fitness for analysing the sentiment across social media text (Pandey, 2018). A positive sentiment is considered to have a compound score of more than 0.05, while negative is less than -0.05. Everything in between is considered neutral. Hence, we have categorized our posts according to these threshold values. All lexical ratings can be found at (Hutto, vaderSentiment, 2020). All scores were analysed against the number of CoVID-19 cases and deaths declared per day in Romania (Portal, 2020) to count for any correlation that could bring an explanation for the daily fluctuations in sentiment scores.

To identify the two main topics of the corpus, we have used LDA (Latent Dirichlet Allocation) modelling from the Python Gensim library (Gensim, 2019). We applied the model on top of the corpus words present in the BOW created previously to make sure we are modelling only the most frequent 5000 words.

Analysis results and interpretations

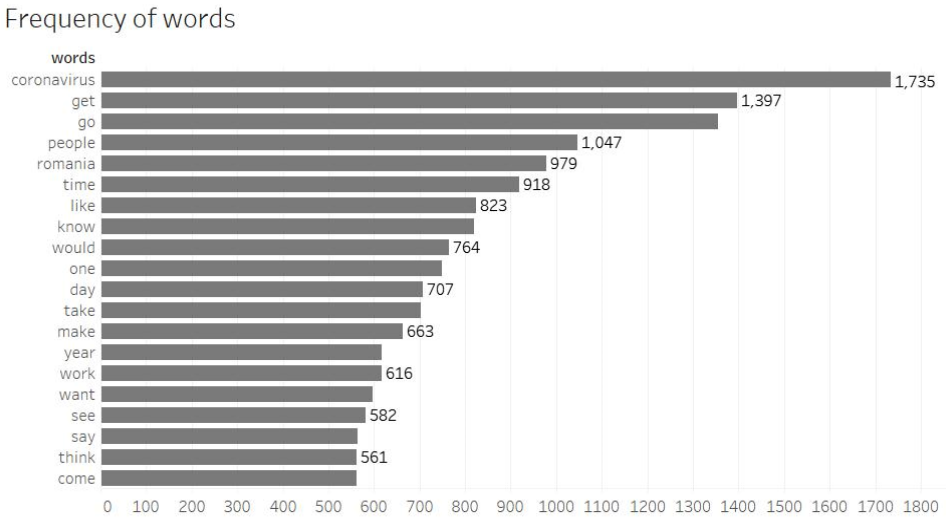
The frequency analysis is displayed in the form of a top-50-words frequency bar chart computed using the Bag of Words for the top 5000 words used in our corpus as displayed in Figure 2. The most frequently used word in the studied subreddit is “coronavirus”, hence people are intensively discussing the situation caused by CoVID-19 or, simply, people are bringing the new coronavirus often in their discussion on Reddit. This finding is aligned with our expectations from the beginning of the study. Along other frequent words we have identified “Romania” and “people” and simple verbs like “know”, “get”, “make“ and “think”. This suggests a common concern for the state of the population of the nation around the pandemic caused by this virus in Romania.

When analysing the number of words per day during the state of emergency (Figure 3), it was easily observed that the numbers were extremely high at the beginning of the period, especially in the first month, causing people to intensively discuss on Reddit. This is explained by a shared concern of the population of the Reddit users around the newly installed situation in the country and the novelty it brings amongst the social spheres affected. In general, the most prominent spikes correspond to important news related to the public

declarations made about CoVID-19 in Romania. One such case is the official proclamation of the city of Suceava being under a restrictive quarantine after an explosion in the number of CoVID-19 infections (Ilie, 2020). This could be perceived as the fact that people were using high quantities of text to express themselves when important news was addressed to the population.

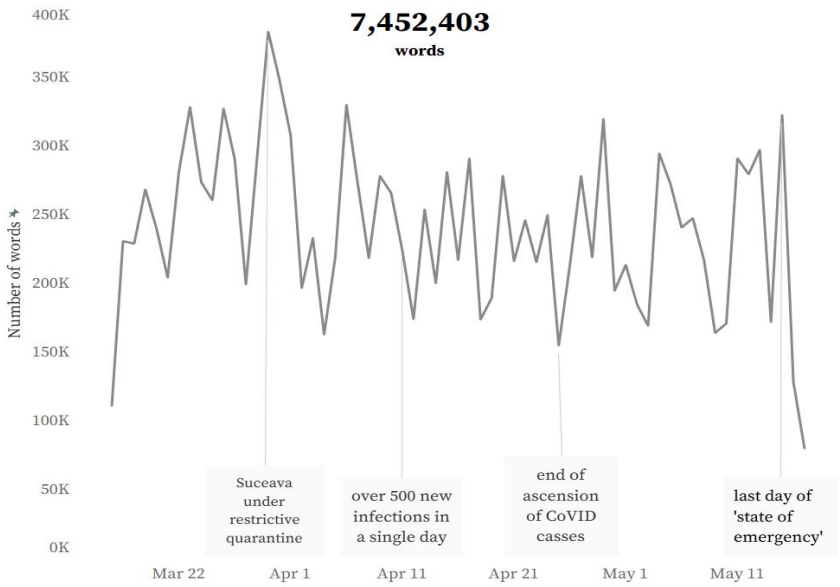
Overall, the compound sentiment scores vary from -0.998, representing the most negative feeling to 1 for the most positive, with a mean of 0.016, hence having a predominantly neutral sentiment across the Reddit posts analysed during the state of emergency in Romania. However, the frequency of posts per sentiment category (Figure 4) reveals a surprising twist from what we anticipated initially: the positive posts have a higher occurrence than the negative ones. This fact might imply that in general, people had a tendency for being more positive when expressing themselves via Reddit posts in Romania during the state of emergency, than they were negative. This in turn, could be interpreted as a generally optimistic attitude in favour of a pessimistic or fearful one, coinciding to what was discovered amongst people in China which are known to be more knowledgeable about the disease and its precautions. (Zhong, et al., 2020)

Figure 2. Frequency bar chart based on the BOW text representations for the 50 most frequent words in corpus – Reddit data from the Romania subreddit retrieved between 16 March 2020 and 16 May 2020



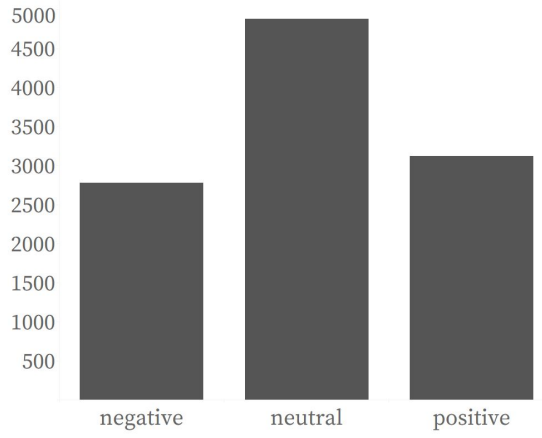
Source: personal computations based on the data from Reddit (2019).

Figure 3. Number of words per day used in Reddit posts and their subsequent comments – corpus data gathered from the Romania subreddit retrieved between 16 March 2020 and 16 May 2020



Source: personal computations based on the data from Reddit (2019).

Figure 4. Frequency of Reddit posts per category of sentiment based on the corpus data gathered from the Romania subreddit retrieved between 16 March 2020 and 16 May 2020



Source: personal computations based on the data from Reddit (2019).

When comparing the normalized mean values for the polarity scores per day, some peaks relate with the increases and decreases in the numbers of deaths and registered incidents of CoVID-19 in Romania (Figure 5). For some increases in the number of declared occurrences of the disease, Reddit more positive posts seem to be recorded, which comes as unexpected from the normal reaction in such situation. This could either be perceived as an outburst of ironical posts, which is easily misinterpreted by the classical sentiment analysis tooling or, simply, as an optimistic attitude in this population in particular which might be more knowledgeable in terms of the virus, as in the study of the knowledge, attitudes, and practices of the Chinese people during the rise of COVID-19 pandemics (Zhong, et al., 2020).

To understand how these variables correlate with each other, we have computed a correlation matrix between the daily mean word length of the posts in the BOW, their daily mean computed Vader score and the number of CoVID-19 caused infections and deaths (Table 3). As shown in the matrix, the infections and deaths have a high positive correlation, which is not surprising as these variables are dependent on each other: the number of infections influences the number of deaths caused by the virus. What is interesting to notice is that the number of words used in the BOW posts are correlating positively with the number of CoVID-19 infections and so does their computed Vader score. Even though the correlations are not high, they can still be significant when compared to the events occurring during the 59 days that were studied here. Going further with the analysis, we have plotted a time series graph to better observe the behaviour of the Vader score in relation to the registered infections and deaths.

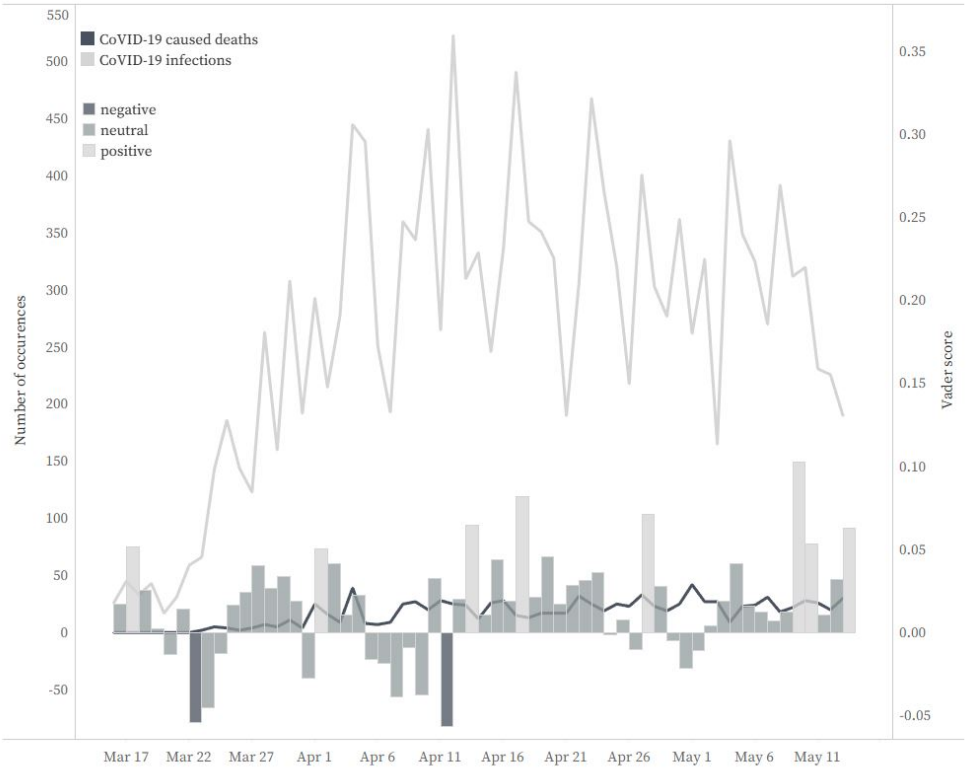
Table 3. Correlation matrix between the posts' in BOW word length, the posts' computed Vader score and CoVID-19 caused deaths and infections as per the statistics provided by the European Centre for Disease Prevention and Control

	Posts' Word Length	**Posts' Vader Score**	**CoVID-19 infections**	**CoVID-19 caused deaths**
Posts' Word Length	1	0.005207	0.269617	-0.0318
Posts' Vader Score	0.005207	1	0.23694	0.18521
CoVID-19 infections	0.269617	0.23694	1	0.613265
CoVID-19 caused deaths	-0.0318	0.18521	0.613265	1

Source: personal computations based on the data from Portal (2020).

During the first weeks of the state of emergency, the analysis reveals that people were expressing more negative feelings than during the latter stages. This could be interpreted as an accommodation of the population with the situation of the country as well as a possible acceptance coming from the Reddit users of Romania. Right before the end of the official state of emergency, people seem to become more enthusiastic as multiple peaks of positive feelings are recorded in the data. This could be easily seen as a positive and optimistic attitude towards the end of the uncomfortable state and the beginning of some more relaxed times. Once again, by analysing the entire set of records for this period, we find out that the general attitude of the population is a neutral one during most days, with 8 days of positive sentiments expressed in the posts of the studied subreddit and only 2 days of slightly negative feelings.

Figure 5. Time series analysis of the evolution of the normalized polarity scores compared to the cases and deaths caused by CoVID-19 - corpus data from the Romania subreddit retrieved between 16 March 2020 and 16 May 2020

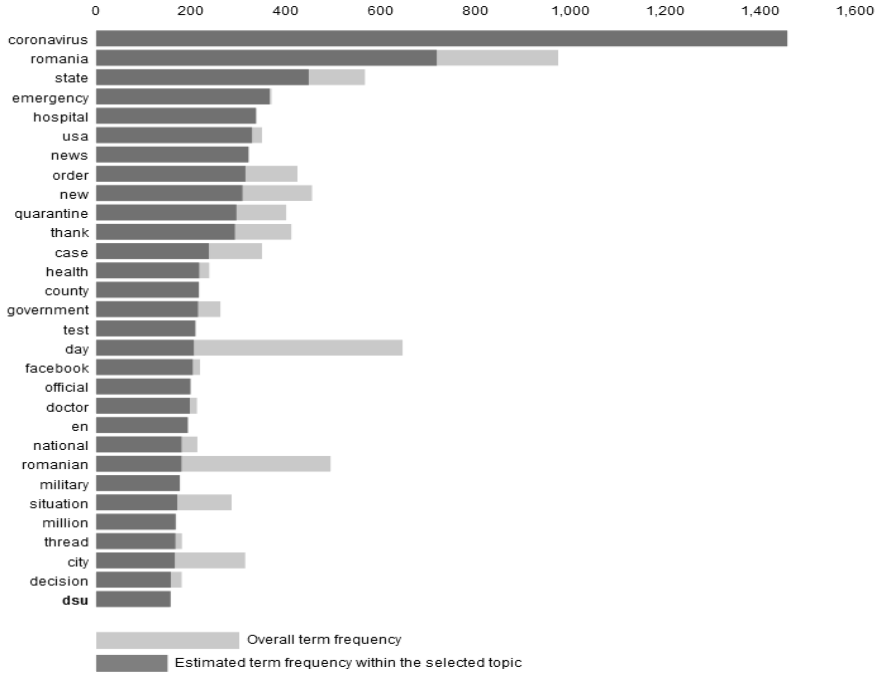


Source: personal computations based on the data from Reddit (2019).

The topic modelling with the LDA algorithm (Figure 6.a) has revealed, for one topic, a primary interest into the topic of Romania's state of emergency caused by CoVID-19: "coronavirus", "Romania", "state" and "emergency" are the top four tokens by relevance in this topic. This comes as a good proof of the fact that the population of the country which usually use Reddit has a keen interest in discussing the latest news both at the country level and at an international level regarding the state of emergency installed. Words like "Bucharest", "Cluj", "national" and "Romania" are pointing us into the direction of discussions of the implications that the virus CoVID-19 has at the national level. On the other side, the term "USA" is identified in the top 10 words used in the modelling of the first topic, as well as the term "news", which again helps us to identify the same interest of discussion. The words "hospital", "quarantine", "case" and "health" are shaping the unanimous concern around the virus' implication in the health system, therefore people are expressing their opinion on the well-being of the country's population in the times of this social crisis.

Figure 6. LDA Topic modelling results for the two most prominent topics discussed in the corpus data gathered from the Romania subreddit retrieved between 16 March 2020 and 16 May 2020

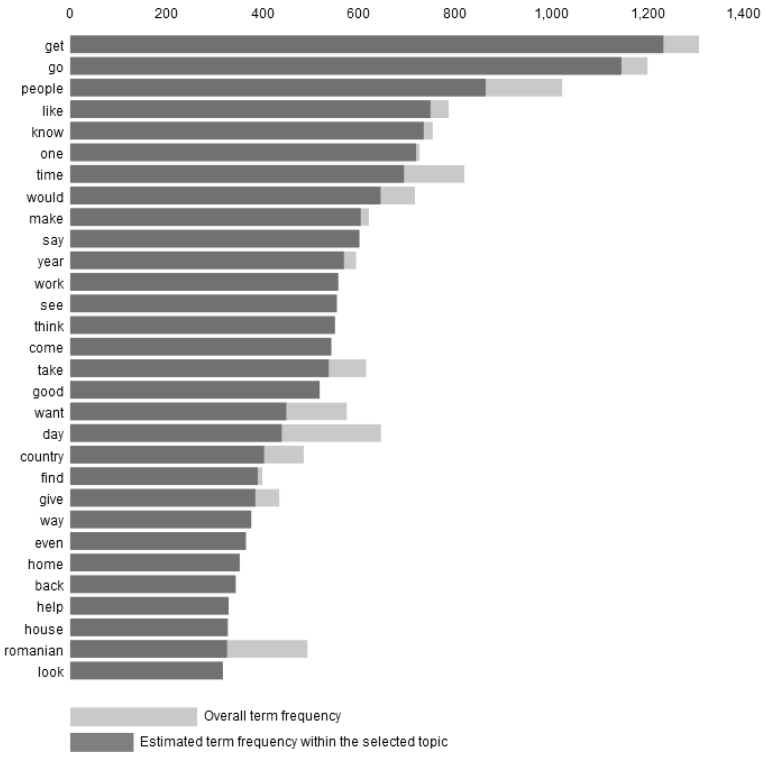
a) Top-30 Most Relevant Terms for Topic 1 (34.7% of tokens)



Source: personal computations based on the data from Reddit (2019).

As for the second topic (Figure 6.b), it seems that people might be discussing actions, such as regular activities in their daily lives: “get”, “go”, “make”, “know” or “work”. The coverage the tokens in this topic is widely spread, with a percentage of 66.5% of the words used. This is because these are commonly used verbs one would prefer when writing non-formal text in general. Tokens like “good” or “like” imply a positive attitude in this topic, which seems to be at the opposite side when compared to the main topic where the dominant feeling would rather be neutral to some words indicating maybe a more negative connotation driven from words like “hospital” and “quarantine”. Moreover, the terms “day”, “home” and “house” present in this topic, bring stronger proof that on the second basis the discussion is focused on the daily living activities of people.

b) Top-30 Most Relevant Terms for Topic 2 (66.5% of tokens)



Source: personal computations based on the data from Reddit (2019).

Discussion

Our research opens a new gate towards the contribution of text mining with social data in the times of a social crisis, with a study case that is focused around the unusual situation caused by the CoVID-2019 pandemics. By analysing what Reddit users have posted during the state of emergency declared in Romania, we have come closer to the understanding of both the overall and daily reaction to this atypical situation our society had to face. Our analysis focused around word and sentiment frequencies, as well as on topic modelling.

First, our initial analysis revealed an extremely high frequency of the word "coronavirus", followed by other terms which implied a general concern around the pandemics and its ramifications. Hence, along this preferred term, "quarantine", "Romanian", "people", "hospital" or "state" were some of the most frequently used words. This brings us close to shaping a global topic that was intensely discussed among Romanians, a topic that was also confirmed by our last stages of the research, the LDA topic modelling. Knowledge was naturally built on top of our findings and a consistent number of conclusions were drawn from only modelling the topic of the Reddit posts.

Going further with the analysis, the overall feeling identified in the corpus of the posts is a neutral to positive one, implying that people were mostly discussing the facts with a few outbursts of more powerful feelings. In this context, we could state the social media represented by our study audience is not necessarily a means to express strong feelings as it is more of an instrument to debate the shortcomings of the pandemics and, probably, its theoretical considerations. However, by analysing the time evolution of the words' frequency in the corpus and the sentiment scores of the posts per day, we could easily derive significant conclusions upon both the tendency to post more in times of important events and the way the feelings were expressed before and after these events.

This research has ultimately come as a contribution to the theoretical findings upon the sociological reaction in times of a crisis as well as the practical implications of using data mining with social media corpus. As per the results of the present analysis, we have taken an insight into what is it that the people in Romania which are using Reddit as a social media platform are discussing about and which is the polarity expressed within the text they are composing in this manner. Our findings revealed that, during a social crisis, in this case caused by the CoVID-19 pandemics, people are actively discussing around the cause of the crisis, the virus itself and around the implications it has upon the well-being of the society. The forum nature of the Reddit platform allowed us to explain why is it that the text discussed and analysed during the lockdown in Romania had an overall neutral to positive feeling. Using the most common text mining techniques has allowed us to understand the social impact the CoVID-19 pandemics with its consequent lockdown had upon Romania's young to middle-aged population that is using Reddit.

We are also rising a few concerns regarding first the generalizability of our dataset as we have used here a limited sample and might not fully represent the population of Romanians who are knowledgeable about CoVID-2019. As (Lazer & Radford, 2017) pointed out, "All of Twitter is a census of Twitter", hence all of Reddit is a census of Reddit. This study can well be continued by analysing what Romanians were discussing on other social media platforms such as Twitter and Facebook to grasp the overall feeling expressed by some different categories of users from these platforms. Second, the ethics of analysing big data in sociology (Herschel & Miori, 2017) should be addressed by both this study as well as any other related one. This means that a certain emphasis should enhance the vulnerabilities of the findings and proposed methodology in the future research. Our aim was to exclusively arrive to a better understanding of how people are facing the serious times of a crisis and not to use these findings for any unethical purpose.

Conclusions and future implications

The present research has resulted into both expected and surprising results: people in Romania have proven to be mainly posting about CoVID-19 issues related to their lives with a general neutral to positive sentiment. In the dynamics generated by the exchange of posts and comments constructed on the Reddit platform, the Romanian users have proven to change their overall expressed sentiments from negative spikes at the beginning of the state of emergency to more positive ones towards the end of this period. This could be interpreted as a general concern among the population of the country that socializes using Reddit around the implications the virus has within the daily social activities of the individuals. When analysed against important events occurring in the country, the polarity scores' fluctuations could be contextualized.

To improve and expand our study for the future we target an application of the present methodology to other popular social networking platforms such as Facebook and Tweeter for a wider targeted audience. In this way, our analysis will address other categories of social media users, which might bring more valuable information to this research. Also, we aim more in-depth sentiment analysis for the Reddit thread specially dedicated to CoVID-19 to identify patterns and tendencies across time. To fully extend this study in this direction, a full analysis of the comments of all analysed posts is desired for a better understanding of how people are reacting to other people's posts. We expect the reactions to be stronger and hence to reveal interesting patterns among the sentiments expressed in Romania's subreddit. To take this research at an international level, we aim for a comparison of the tendencies identified in Romania with the posts people have discussed worldwide.

ABBREVIATIONS:

BOW	Bag of Words
LDA	Latent Dirichlet Allocation
CoVID-19	Coronavirus Disease 2019

REFERENCES

Ampofo, L., Collister, S., O'Loughlin, B., & Chadwick, A. (2015). Text Mining and Social Media: when Quantitative Meets Qualitative and Software Meets People. In P. Halfpenny, R. Procter, P. Halfpenny, & R., Procter (Eds.), Innovations in Digital Research Methods (pp. 161-191). SAGE.

Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., & Vorontsov, K. (2016). Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. *Advances in Computational Intelligence*, 169-184. Doi:https://doi.org/10.1007/978-3-319-62434-1_14.

Borovikova, E. (2011, November 18). sklearn.feature_extraction.text.CountVectorizer. Retrieved May 31, 2020, from Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

Cioban, Ş., & Vîntoiu, D. (2020, May 20). Covid19_sentiment_analysis. Retrieved from https://github.com/stefanacioban/covid19_sentiment_analysis.

Dawn Breslin, S., Enggaard, T., Blok, A., Gårdhus, T., & Pedersen, M. (2020, May 23). How We Tweet About Coronavirus, and Why: A Computational Anthropological Mapping of Political Attention on Danish Twitter during the COVID-19 Pandemic. Science, Medicine, and Anthropology. Retrieved June 7, 2020, from http://somatosphere.net/forumpost/covid19-danish-twitter-computational-map/.

De Vos, J. (2020, May). The effect of COVID-19 and subsequent social distancing on travel behavior. *Transportation Research Interdisciplinary Perspectives*, 5. Doi:https://doi.org/10.1016/j.trip.2020.100121.

Deepa, Y., Manjunath, T., & Ravindra, H. (2019). Review on Natural Language Processing Trends and Techniques Using NLTK. In Recent Trends in Image Processing and Pattern Recognition (pp. 589-606). Solapur, India: Springer. Doi:10.1007/978-981-13-9187-3_53.

Dingwall, R., Hoffman, L., & Staniland, K. (2013, February). Introduction: Why a Sociology of Pandemics? In R. Dingwall, L.M. Hoffman, & K. Staniland, Pandemics and Emerging Infectious Diseases: The Sociological Agenda (Vol. 35, pp. 167-173). Wiley-Blackwell. Doi:https://doi.org/10.1111/1467-9566.12019.

Fersini, E. (2017). Chapter 6 - Sentiment Analysis in Social Networks: A Machine Learning Perspective. In F. Pozzi, E. Fersini, E. Messina, & B. Liu, Sentiment Analysis in Social Networks (pp. 91-111). Milan, Italy: Morgan Kaufmann. Doi:https://doi.org/10.1016/B978-0-12-804412-4.00006-1.

Gaspar, R., Pedro, C., Panagiotopoulos, P., & Seibt, B. (2016). Beyond Positive or Negative: Qualitative Sentiment Analysis of Social Media Reactions to Unexpected Stressful Events. *Computers in Human Behavior*, 56, 179-191.

Gensim. (2019, November 1). models.ldamodel – Latent Dirichlet Allocation. Retrieved May 31, 2020, from Gensim: https://radimrehurek.com/gensim/models/ldamodel.html.

Habermas, J. (1991). The Public Sphere. In C. Mukerji, & M. Schudson, Rethinking Popular Culture: Contemporary Perspectives in Cultural Studies (pp. 389-404). Berkeley, California, USA: University of California Press.

Herschel, R., & Miori, V. (2017, May). Ethics & Big Data. *Technology in Society*, 49, 31-36. Doi:https://doi.org/10.1016/j.techsoc.2017.03.003.

Hutto, C. (2020, May 20). vaderSentiment. vader_lexicon.txt. Atlanta, GA. Retrieved May 31, 2020, from https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt.

Hutto, C., & Gilbert, E. (2014, May 16). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International AAAI Conference on Weblogs and Social Media. AAAI Publications. Retrieved May 31, 2020, from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

Ilie, L. (2020, March 31). Romania places eastern city of Suceava under quarantine. (R. Gopalakrishnan, Ed.) Retrieved June 22, 2020, from Reuters: https://www.reuters.com/article/us-health-coronavirus-romania/romania-places-eastern-city-of-suceava-under-quarantine-idUSKBN21I0KC.

Kasapoglu, A., & Akbal, A. (2020, April 25). Relational Sociological Analysis of Uncertainties: The case of COVID-19 In Turkey. *Advances in Social Sciences Research Journal*, 7(4), 197-228.

Krippendorff, K. (2005, January). The Social Construction of Public Opinion. In E. Wienand, J. Westerbarkey, A. Scholl, E. Wienand, J. Westerbarkey, & A. Scholl (Eds.), Kommunikation über Kommunikation. Theorie, Methoden und Praxis. Festschrift für Klaus Merten. (p. 130). Wiesbaden, Hesse, Germany: VS-Verlag. Doi:10.1007/978-3-322-80821-9_10.

Lakoff, G., & Johnson, M. (2003). Metaphors We Live By (first edition ed.). Chicago, Illinois, USA: University of Chicago Press.

Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology,* 19-39. Doi:https://www.annualreviews.org/doi/10.1146/annurev-soc-060116-053457.

Mediakix. (2018, December 28). The Top 8 Reddit Statistics On Users, Demographics & More. Retrieved May 31, 2020, from https://mediakix.com/blog/reddit-statistics-users-demographics/.

Medvedev, A., Lambiotte, R., & Delvenne, J.-C. (2018, October 25). The anatomy of Reddit: An overview of academic research. Dynamics On and of Complex Networks III. Doi:10.1007/978-3-030-14683-2_9.

Mutzel, S. (2015). Facing Big Data: Makin sociology relevant. *Big Data & Society*, 2(2). Doi:https://doi.org/10.1177/2053951715599179.

Ornell, F., Schuch, J., Sordi, A., & Kessler, F. (2020). "Pandemic fear" and COVID-19: mental health burden and strategies. *Brazilian Journal of Psychiatry*, 42(3). Doi: http://dx.doi.org/10.1590/1516-4446-2020-0008.

Pandey, P. (2018, September 23). Simplifying Sentiment Analysis using VADER in Python (on Social Media Text). Analytics Vidhya. Retrieved May 31, 2020, from https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f.

Portal, E.O. (2020, May 13). COVID-19 cases worldwide. COVID-19 Coronavirus data. ECDC European Centre for Disease Prevention and Control. Doi:10.2906/101099100099/1.

Python. (2014, February 22). Welcome to Python. Retrieved from https://www.python.org/.

Reddit. (2020, January 10). Reddit Privacy Policy. Retrieved May 31, 2020, from Reddit: https://www.redditinc.com/policies/privacy-policy.

Reddit. (2018, September 11). reddit.com: api documentation. Retrieved from Reddit: https://www.reddit.com/dev/api.

Reddit. (2019, January 2). Reddit Romania. Retrieved May 31, 2020, from Reddit: https://www.reddit.com/r/Romania/.

Reddit. (2019, Nov 5). Reddit- the front page of the internet. Retrieved May 31, 2020, from Reddit: https://www.reddit.com/.

Ristoski, P., & Paulheim, H. (2016, January). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics*, 36, 1-22. Doi:https://doi.org/10.1016/j.websem.2016.01.001.

Shigemura, J., Ursano, R., Morganstein, J., Kurosawa, M., & Benedek, D. (2020, February). Public Responses to the Novel 2019 Coronavirus (2019-nCoV) in Japan: mental health consequences and target populations. Psychiatry Clin Neurosci. Doi: https://doi.org/10.1111/pcn.12988.

SimilarWeb. (2020, April 20). Analytics - Market Share Stats en.reddit.com. Retrieved from SimilarWeb: https://www.similarweb.com/website/reddit.com#display.

Subramanian, D. (2019, August 22). Text Mining in Python: Steps and Examples. Retrieved May 31, 2020, from https://medium.com/towards-artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b.

Swasey, C., Winter, E., & Sheyman, I. (2020). The Staggering Economic Impact of the Coronavirus Pandemic. Data for Progress. Retrieved May 31, 2020, from https://www.dataforprogress.org/memos/coronavirus-economic-impact.

Tableau. (2009, September 27). Tableau. Retrieved from https://www.tableau.com/.

Usai, A., Pironti, M., Mital, M., & Mejri, C. (2018, October 10). Knowledge discovery out of text data: a systematic review via text mining. *Journal of Knowledge Management,* Emeral Publishing Limited. Doi: http://dx.doi.org/10.1108/JKM-11-2017-0517.

Yandex Technologies. (2019, July 24). Yandex. Retrieved May 31, 2020, from About machine translation: https://tech.yandex.com/translate/.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010, December 1). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics,* 1, 43-52. Doi:10.1007/s13042-010-0001-0.

Zhong, B.-L., Luo, W., Li, H.-M., Zhang, Q.-Q., Liu, X.-G., Li, W.-T., & Li, Y. (2020). Knowledge, attitudes, and practices towards COVID-19 among Chinese residents during the rapid rise period of the COVID-19 outbreak: a quick online cross-sectional survey. *International Journal of Biological Sciences*, 16(10), 1745-1752. Doi:10.7150/ijbs.45221.