Em Questão
ISSN: 1807-8893
ISSN: 1808-5245
emquestao@ufrgs.br
Universidade Federal do Rio Grande do Sul
Brasil

COVID-19 and the circulation information on social networks: analysis in a Brazilian Facebook group about the Coronavirus

Farias Cordeiro, Douglas; Souza Rocha, Anelise; Machado Vieira, Larissa; Kelvis Cassiano, Kátia; Rosa Da Silva, Núbia

COVID-19 and the circulation information on social networks: analysis in a Brazilian Facebook group about the Coronavirus

Em Questão, vol. 27, núm. 3, 2021

Universidade Federal do Rio Grande do Sul, Brasil

Disponible en: https://www.redalyc.org/articulo.oa?id=465668004003

DOI: https://doi.org/10.19132/1808-5245273.42-67

COVID-19 and the circulation information on social networks: analysis in a Brazilian Facebook group about the Coronavirus

Douglas Farias Cordeiro 1
Universidade Federal de Goiás, Goiânia, Brasil
cordeiro@ufg.br

Anelise Souza Rocha 2
Universidade Federal de Goiás, Goiânia, Brasil
anelisesrocha@gmail.com

Larissa Machado Vieira 3
Universidade Federal de Goiás, Goiânia, Brasil
vieira.mlarissa@gmail.com

Kátia Kelvis Cassiano 4
Universidade Federal de Goiás, Goiânia, Brasil
katiakelvis@ufg.br

Núbia Rosa Da Silva 5
Universidade Federal de Catalão, Catalão, Brasil
nubia@ufcat.edu.br

DOI: https://doi.org/10.19132/1808-5245273.42-67
Redalyc: https://www.redalyc.org/articulo.oa?id=465668004003

Recepción: 20 Agosto 2020
Aprobación: 09 Febrero 2021

Abstract:

This article aims to quantify and qualify the information circulating in social media groups about COVID-19, the subjects covered in posts, as well as the possible relations with other subjects, events or social events, in order to generate a representative panorama of perception and social reaction to the coronavirus pandemic. For this, statistical techniques, data mining and machine learning are used to the characterization, pattern detection, and grouping of textual data. The experiments are carried out on a dataset of textual data extracted from a Brazilian public group about COVID-19 (SARS-cov-2) of the social network Facebook. Statistical analyzes are crossed with data on the advance of the number of infected, and with specific political-social events, revealing variations and influences in terms of participation and engagement in the analyzed group. In addition, through the results obtained by the clustering method used, two main groups of posts are detected, the first presenting a content pattern geared to governmental issues, and the second to personal issues. The results achieved still allow a reflection on the possible social impacts of the creation or absence of public policies to deal with the COVID-19 pandemic.

Keywords: Covid-19, SARS-cov-2, Online social networks, Data mining, Descriptive analysis.

Notas de autor

[1] Doutor; Universidade Federal de Goiás, Goiânia, GO, Brasil
cordeiro@ufg.br

[2] Mestranda; Universidade Federal de Goiás, Goiânia, GO, Brasil
anelisesrocha@gmail.com

[3] Mestre; Universidade Federal de Goiás, Goiânia, GO, Brasil
vieira.mlarissa@gmail.com

[4] Doutora; Universidade Federal de Goiás, Goiânia, GO, Brasil
katiakelvis@ufg.br

[5] Doutora; Universidade Federal de Catalão, Catalão, GO, Brasil
nubia@ufcat.edu.br

Resumo:

Este artigo tem como objetivo quantificar e qualificar as informações que circulam nas redes sociais sobre o COVID-19, os assuntos abordados nas postagens, bem como as possíveis relações com outros assuntos, eventos ou eventos sociais, de forma a gerar um panorama representativo da percepção e reação social à pandemia de coronavírus. Para isso, técnicas estatísticas, mineração de dados e aprendizado de máquina são utilizadas para a caracterização, detecção de padrões e agrupamento de dados textuais. Os experimentos são realizados em um conjunto de dados textuais extraídos de um grupo público brasileiro sobre o COVID-19 da rede social Facebook. As análises estatísticas são cruzadas com dados sobre o avanço do número de infectados e com eventos político-sociais específicos, revelando variações e influências em termos de participação e engajamento no grupo analisado. Além disso, através dos resultados obtidos pelo método de agrupamento utilizado, são detectados dois grupos principais de postagens, o primeiro apresentando um padrão de conteúdo voltado para questões governamentais e o segundo para questões pessoais. Os resultados alcançados permitem ainda uma reflexão sobre os possíveis impactos sociais da criação ou ausência de políticas públicas para o enfrentamento da pandemia COVID-19.

Palavras-chave: Covid-19, Redes sociais digitais, Grupos sociais, Mineração de dados, Análise descritiva.

1 Introduction

In December 2019, the city of Wuhan, China, witnessed the emergence of one of the most alarming pandemics recorded in the Contemporary Age, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-cov-2), known as COVID-19. Identified as a zoonotic coronavirus, similar to other known viruses, such as the Severe Acute Respiratory Syndrome (SARS-cov) and the Middle East Respiratory Syndrome Coronavirus (MERS-cov) (LIU et al., 2020), COVID-19 overtook in early April 2020, the number of one million infected around the world, with more than fifty thousand deaths registered, according to data released by the World Health Organization (WHO, 2020a). According to information presented in Wu and McGoogan (2020), the three aforementioned viruses have considerably similar characteristics, with presentation of fever, cough and problems in the lower respiratory tract, with attenuation associated with age, as well as underlying conditions of other diseases. However, COVID-19 was potentially more aggressive in epidemiological terms, with an exponential increase in the number of infected people.

According to WHO (2020a), the number of coronavirus cases in Italy, for example, jumped from 888 on February 29, 2020, to 105,792 on March 31, 2020. In a study by Pan et al. (2020), 21 patients with pneumological infection caused by COVID-19 were followed, and it was found that the hospitalization period varies within the range of 11 to 26 days, which, combined with the growing number of infected people, ends up leading health systems to collapse. In Brazil, the first case of contamination by coronavirus was confirmed on February 26, 2020 (MELO et al., 2020), and on March 31, 2020, 5,717 infections and 201 deaths were recorded (WHO, 2020a). In a global context, the number of infected people increases exponentially from the moment the first infections occur.

Based on the established pandemic scenario, declared by the WHO (2020b) on 11 March, the establishment of policies and containment strategies that contemplate social isolation, quarantine and border closure was adopted by several countries. At the same time, due mainly to the large number of isolated people in their homes, a significant increase in the use of the Internet for information and entertainment purposes was observed. A survey conducted by the German portal Statista, between 16 and 20 March, with a sample of 12,845 individuals aged between 16 and 64 years, distributed in different countries around the world, found a 40% growth in the use of laptops, and of 70% in the use of smartphones, among which there was a 44% growth in the use of social networks.

It is interesting to note that the possibilities of interaction promoted by the popularization of mobile devices connected to the Internet, considering the context of Web 2.0 (O'REILLY, 2005), enabled an emergence of communities and virtual groups, in which individuals come together to seek knowledge, disseminate information, and discuss different topics. Such spaces are characterized by the continuous sharing of innumerable contents, under different presentation formats (text, image and video), and new forms of relationship and approaches on common interests are built (PENNI, 2017).

Considering the characteristic of the natural collectivity of Web 2.0, even though there are barriers regarding the establishment of direct interactions between two individuals on the network, the accessibility of a given content, generated or shared, is notably possible and expected. This phenomenon clearly occurs in the context of social networks, where there is a certain freedom of expression, and the identity of individuals are not verified, which ends up promoting greater participation and openness in terms of opinions, feelings, debates, or dissemination of information (CERCEL; TRAUSAN-MATU, 2014).

The environment promoted by social networks stimulates interactivity and relationships between generators and consumers of information. This interaction promotes a rupture of the barriers traditionally defined between these elements, being essential for the foundation of the circulation of information (SREEJESH et al., 2020). Furthermore, the emergence of new forms of communication give rise to scenarios where information emerges collaborative or even misinformation (LOGAN, 2016).

Based on this, this article aims to carry out an analysis on the textual data of a set of posts published in a Brazilian group on Coronavirus on the social network Facebook, from January to March 2020. The central focus is quantify and qualify the circulating information in this group, the themes dealt with, as well as the possible relations with other subjects, generating a representative panorama of the perception and social reaction in face of the COVID-19 pandemic.

The analysis carried out still seek to generate inputs for reflections that make it possible to relate the possible relationships and influences of specific events on the effectiveness of the users' participation in the analyzed social group. For that, statistical techniques will be used for descriptive analysis of the data, in order to reveal the implicit patterns of the collected data, as well as methods of analysis of similarity between terms and between sentences named Doc2Vec, based on data mining and machine learning, in order to infer groups of posts that indicate the subjects covered. In addition, the development of the experiments follows a systematization of activities, which is interesting from the point of view of data analysis on social networks. The analysis carried out follows a paradigm based on Big Data and algorithms, considering individuals in a performative way, from the data extracted from the social network, as described in Fisher and Mehozay (2019).

In December 2019, the city of Wuhan, China, witnessed the emergence of one of the most alarming pandemics recorded in the Contemporary Age, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-cov-2), known as COVID-19. Identified as a zoonotic coronavirus, similar to other known viruses, such as the Severe Acute Respiratory Syndrome (SARS-cov) and the Middle East Respiratory Syndrome Coronavirus (MERS-cov) (LIU et al., 2020), COVID-19 overtook in early April 2020, the number of one million infected around the world, with more than fifty thousand deaths registered, according to data released by the World Health Organization (WHO, 2020a). According to information presented in Wu and McGoogan (2020), the three aforementioned viruses have considerably similar characteristics, with presentation of fever, cough and problems in the lower respiratory tract, with attenuation associated with age, as well as underlying conditions of other diseases. However, COVID-19 was potentially more aggressive in epidemiological terms, with an exponential increase in the number of infected people.

According to WHO (2020a), the number of coronavirus cases in Italy, for example, jumped from 888 on February 29, 2020, to 105,792 on March 31, 2020. In a study by Pan et al. (2020), 21 patients with pneumological infection caused by COVID-19 were followed, and it was found that the hospitalization period varies within the range of 11 to 26 days, which, combined with the growing number of infected people, ends up leading health systems to collapse. In Brazil, the first case of contamination by coronavirus was confirmed on February 26, 2020 (MELO et al., 2020), and on March 31, 2020, 5,717 infections and 201 deaths were recorded (WHO, 2020a). In a global context, the number of infected people increases exponentially from the moment the first infections occur.

Based on the established pandemic scenario, declared by the WHO (2020b) on 11 March, the establishment of policies and containment strategies that contemplate social isolation, quarantine and border closure was adopted by several countries. At the same time, due mainly to the large number of isolated people in their homes, a significant increase in the use of the Internet for information and entertainment purposes was observed. A survey conducted by the German portal Statista, between 16 and 20 March, with a sample of 12,845 individuals aged between 16 and 64 years, distributed in different countries around the world, found a 40% growth in the use of laptops, and of 70% in the use of smartphones, among which there was a 44% growth in the use of social networks.

It is interesting to note that the possibilities of interaction promoted by the popularization of mobile devices connected to the Internet, considering the context of Web 2.0 (O'REILLY, 2005), enabled an emergence of communities and virtual groups, in which individuals come together to seek knowledge, disseminate information, and discuss different topics. Such spaces are characterized by the continuous sharing of innumerable contents, under different presentation formats (text, image and video), and new forms of relationship and approaches on common interests are built (PENNI, 2017).

Considering the characteristic of the natural collectivity of Web 2.0, even though there are barriers regarding the establishment of direct interactions between two individuals on the network, the accessibility of a given content, generated or shared, is notably possible and expected. This phenomenon clearly occurs in the context of social networks, where there is a certain freedom of expression, and the identity of individuals are not verified, which ends up promoting greater participation and openness in terms of opinions, feelings, debates, or dissemination of information (CERCEL; TRAUSAN-MATU, 2014).

The environment promoted by social networks stimulates interactivity and relationships between generators and consumers of information. This interaction promotes a rupture of the barriers traditionally defined between these elements, being essential for the foundation of the circulation of information (SREEJESH et al., 2020). Furthermore, the emergence of new forms of communication give rise to scenarios where information emerges collaborative or even misinformation (LOGAN, 2016).

Based on this, this article aims to carry out an analysis on the textual data of a set of posts published in a Brazilian group on Coronavirus on the social network Facebook, from January to March 2020. The central focus is quantify and qualify the circulating information in this group, the themes dealt with, as well as the possible relations with other subjects, generating a representative panorama of the perception and social reaction in face of the COVID-19 pandemic.

The analysis carried out still seek to generate inputs for reflections that make it possible to relate the possible relationships and influences of specific events on the effectiveness of the users' participation in the analyzed social group. For that, statistical techniques will be used for descriptive analysis of the data, in order to reveal the implicit patterns of the collected data, as well as methods of analysis of similarity between terms and between sentences named Doc2Vec, based on data mining and machine learning, in order to infer groups of posts that indicate the subjects covered. In addition, the development of the experiments follows a systematization of activities, which is interesting from the point of view of data analysis on social networks. The analysis carried out follows a paradigm based on Big Data and algorithms, considering individuals in a performative way, from the data extracted from the social network, as described in Fisher and Mehozay (2019).

2 Online social networks and data analysis

The exponential growth of the Internet since the 1990s, attributed to technological development, has instituted new forms of production, dissemination and sharing of information. Social media is a class of information technologies that support interpersonal communication and collaboration through Internet-based platforms, which provides the environment for the formation of dynamic structures for connecting people and interacting - social networks (KANE et al., 2014). The social networks are a set of interrelated nodes that establish bonds generally conceptualized as a social relationship ("friend of" or "boss of") or a dyadic interaction ("conversation with", "sells to", "works with").

It is interesting to note that the advent of Web 2.0 represents an important milestone in the evolution of communications, fostering an environment of media convergence and circulation of information in virtual spaces, as well as an amplification in the access and manipulation of data, both by the existing technologies, as well as advances in the availability and access models. In this context, every individual connected to the Internet ends up becoming a potential content generator, whatever the connected purposes. Belk (2014) state the technologies that identifies the Web 2.0 have allowed and enhanced the development of environments and sharing spaces. Considering the character of the intrinsic collectivity of Web 2.0 and in view of the impossibility of direct interactions between two individuals on the network, a generated or shared content may be accessible, since the free flow of data is natural to virtual spaces. This phenomenon can be clearly seen in the context of social networks, which, as punctuated by Lipschultz (2018), stand out for their freedom of expression, and for the non-mandatory identification of the individual, providing greater delivery by their users within relation to sharing opinions, feelings, or even in discussions on the network.

In addition, Web 2.0 is also reflected in the productive structure of information in virtual spaces, through the emergence of new forms of communication that are decentralized, personalized and interactive, which, enhanced by the wide and democratic access to technological devices (MCLUHAN, 1994), end up generating scenarios where the figures of citizen journalism, collaborative information, or even disinformation emerge.

The context of Web 2.0 and the possibilities of interaction promoted by the popularization of mobile devices connected to the Internet leveraged important changes in the ways of communicating. It is possible to observe the emergence of virtual communities (LEVY, 1997), where individuals come together to discuss different topics. In these spaces, a number of contents are shared and new ways of relating and grouping around common subjects are established. On the other hand, it is also important to note that the democratization of internet access contrasts with a continuous movement of individualization, deterritorialization and inequality, promoted by content personalization, automated by the evolution of suggestion and automated selection algorithms (JUST; LATZER, 2017).

In Brazil, 70% of Brazilians (about 150 million people) have access to the Internet and, of these, 96.2% (about 144 million people) are active on social networks (KEMP, 2020). The most accessed platforms are mainly YouTube, Facebook, Whatsapp, Instagram, Twitter and Linkedin, in that order, with an increase of 8.2% in social network users in Brazil was reported between April 2019 and January 2020 (KEMP, 2020). In addition, between January 2020 and March 2020, there was a 50% increase in social media usage time by Brazilians (STATISTA, 2020a).

Social networks have become a field of study for research related to the organization and treatment of large amounts of data, in addition to providing an ideal environment for extracting knowledge through the application of data mining techniques. The most common structuring elements of social networks such as user profiles, comments, updates, evaluations and metadata are often used as data sources. Through profiles, for example, it is possible to identify people with common interests and map the relationships (ARNABOLDI et al., 2017; LI; DAS, 2020), or to check rumors, misinformation and fake news (FERRARA, 2017; HUSSAIN et al., 2018; AHSAN; KUMARI; SHARMA, 2019).

Social networks generally exhibit a rich internal structure, in which users, through their activity and involvement, define different types and intensity of interactions. As potential analysis in social networks, Tang et al. (2009) quantitatively analyzed social influence at the level of a given content, identifying representative nodes (users) in the context of the topics and the connections created from the degree of influence. The study demonstrated that some connections are characterized by high bandwidth and diversity of context, exhibiting high efficiency of information diffusion. Thus, it was possible to identify network connections where higher quality interactions occur and how much influence in a given context.

Considering the relevance regarding propagation and social networks as a means of quick exchange of information when disasters occur, some studies (KEIM; NOJI, 2011; KRYVASHEYEU et al., 2016; KIM; BAE; HASTAK, 2018; DRAGOVIĆ et al., 2019; LIU; ZHANG; ZHANG, 2020) used data extraction techniques, especially on Facebook and Twitter, to analyze the information circulating on social networks in order to map human behavior and understand its impact during emergency situations, such as natural disasters, epidemics, terrorist attacks and public order demonstrations.

Understanding social networks as the new forum for collective intelligence, social convergence and community activism, in Keim and Noji (2011) is discussed the immediate consequences of the 2010 earthquake in Haiti, with regard to the information circulating in discussion groups on MySpace and Facebook. The first finding presented in their work is that much of what people around the world were learning about the earthquake came from these sources. In addition to sharing information, these focus groups were also used for donations and to offer comfort and support (psychological benefits) to vulnerable populations. Such a functional organization observed in social networks in the context of that disaster, opposes the ingrained vision of social media for transmitting unidirectional information such as radio, TV, newspapers and magazines.

The study suggests, conclusively, that large-scale interaction on social networks can alter the way in which the world reacts to disasters, as the positive effects of this response would increase people's degree of resilience, personal and collective responsibility, decreasing risks in socioeconomic reorganization. Despite the potentialities highlighted, the authors emphasize the importance of managing information on social networks, in order to prevent the spread of rumors that could lead to widespread panic and negative social impacts.

Considering that the dissemination of information on social networks provides the situational awareness of its users, in Kryvasheyeu et al. (2016) is used the spatio-temporal distribution of messages related to disasters to suggest a model for real-time monitoring and evaluation of the disaster itself. The authors treated Twitter activity as a case study before, during and after Hurricane Sandy, pointing out in their results that real and perceived threats by people, together with the effects of physical disasters, are directly observable through the intensity and composition of the flow from Twitter messages to a wide range of disasters. That user activity on Twitter is strongly correlated with economic damage per capita when inflicted by the hurricane. The authors suggest the use of data from social networks in the rapid assessment of the damage caused by a large-scale disaster.

From the perspective of mental health, disasters in general have substantial social consequences. Analyzing data from Twitter, in Gruebner et al. (2018) are identified emotions in space-time relations – an increased discomfort, that is, negative emotions accumulated after a disaster, compared to the emotions presented during the disaster. The study suggests that significant associations of negative emotional responses in the space and time of a natural disaster can be used in alert systems, in the identification of regions or social groups that need attention with a view to mental health.

An important question addressed by Liu et al. (2020a) is that despite the opportunities offered by social networks in communicating disasters compared to traditional media, freedom of opinion can result in distorted information. To assess this impact, the authors analyzed the functional structure of social networks, according to the ability to control and influence the nodes, in communicating information about disasters and in communicating the risk of disasters or threats. The results showed that the activity of nodes in small groups favors the dissemination over long distances, while the communication of disaster risks is strongly dependent on the activity of key nodes (opinion leaders) and, furthermore, this performance is prone to generation of rumors.

3 Methodology

The knowledge discovery from databases is something of great importance and interest. In an environment dominated by Big Data, it demands the use of automated and intelligent techniques, which allow the generation of strategic information for the purposes related to the problem to be addressed. It is essential to execute a series of steps to guarantee the assertiveness of the results to be obtained, minimizing the impacts induced by noise in the data sets.

The experiments proposed in this work are guided by the process called Knowledge Discovery in Databases (KDD), proposed by Fayyad, Piatetsky-Shapiro and Smith (1996). KDD is a process that guides the generation of information and the recognition of patterns based on the execution of five steps: selection, pre-processing, transformation, data mining and interpretation. One of the main advantages of KDD is the fact that it is an interactive process, as it is presented in a sequential and organized way and iterative because it allows interventions in its activities (PROVOST; FAWCETT, 2013).

The first step to be performed is the selection of the data. During this stage, the databases to be used to solve the problem are defined, as well as the actions related to the collection of this data Han, Kamber and Pei (2016). The data used comes from a Brazilian public group from the online social network Facebook. Facebook was launched in early 2004, reached approximately two billion active users in the world in 2020 and 120 million active users in Brazil (STATISTA, 2020b). In the context of the data to be explored in our work, the group selected for the experiments proposes the dissemination of information about Coronavirus, having its first post published on January 25, 2020. The group has approximately 45 thousand active users.

During the selection phase, after defining the data sample, the first activity is the data extraction. For this, a Web Scraping solution (JARMUL; LAWSON, 2017) based on the Python programming language was built. All posts from the defined time period were extracted, from January 2020 to March 2020, including the following attributes: post id, user id, post content, post data, number of likes, and number of comments. The identification data of users were neither collected nor stored, and the user id code was changed at run time by the application of the unidirectional cryptographic dispersion function Message-Digest algorithm 5 (MD5) (RIVEST, 1992), without allowing the reconstruction of the original values. The 7,523 posts published were extracted, of which 68% refer to posts containing textual data and 32% refer to posts containing only image or video data. For the experiments, only posts with textual data were considered.

After selecting the data, the pre-processing step must be performed, where data cleaning activities are performed, treating possible records that present noise or missing data, in order to guarantee the quality of the analyzes performed (HAN et al., 2016). During the pre-processing phase, routines were applied to clean noise and remove unnecessary data for analysis purposes. We use rules based on regular expressions to remove links, non-alphabetic characters, quotes from other users (which are specified by terms starting with the @ symbol), hashtags (identified by the # symbol), stop words and the use of repeated letters (example: "I likeeee this") (SALLOUM; AL-EMRAN; SHAALAN, 2017; LI et al., 2019). The occurrences of emoticons were identified through regular expressions and treated by replacing them with the corresponding words (WANG; CASTANON, 2015).

The third step of the process refers to the transformation of the data into a format more suitable for the purposes of analysis. After the execution of the treatment routines, the data were transformed in two formats, based on the defined analysis purposes. First, a database structured in a csv dataset (comma separated values) was generated, containing all the attributes of the extracted posts, which is used for the development of the descriptive analysis. In addition, a textual corpus was also generated, containing only the post id and text data, which is used by similarity detection, grouping and content analysis methods.

Under the KDD methodology, data mining is considered the most important step, since it is mainly responsible for generating information. Data mining, according to Tan et al. (2019) refers to a set of processes aimed at the exploration and analysis of large data sets, with the purpose of raising standards, associations, anomalies, and forecasts. Provost and Fawcett (2013) highlight that data mining corresponds to the application of computational solutions that allow extracting knowledge from pre-processed databases. In general, the features of data mining can be explored under two approaches: descriptive and predictive. Descriptive methods are aimed at characterizing, summarizing and discriminating data, while predictive methods seek inference or prediction from analysis in databases Han et al. (2016).

The methods of descriptive analysis applied are based on statistical, computational and information visualization techniques, aimed at surveying unidentified trends and patterns (GREENELTCH, 2019). The pre-processed data are used for the generation of graphs that contemplate the temporal evolution in the number of posts, the most frequently used terms, and the engagement of the posts, which refers to the involvement obtained in the posts by the users, given by the sum of the number of likes and comments.

A point of fundamental importance refers to the fact that in the natural language processing routines, the information generated depends not only on the terms present in isolation, but also on the context in which they are found, allowing the discovery of patterns and groups specific textual content. One of the objectives of this article is to find the possible groups of posts present in the selected sample, in order to determine the content to which these groups are dealing. For this, we use the Natural Language Processing (NLP) technique called Doc2Vec (LE; MIKOLOV, 2014).

The Paragraph Vector method, also called Doc2Vec, was introduced by Mikolov et al. (2013) and can be described as a NLP tool for the representation of documents, being considered a generalization of the Word2Vec method (MIKOLOV et al., 2013). In general, this technique can be described as an unsupervised learning model, which is based on distributed vector representations of the terms or words of a text. From this, the texts referring to the considered database can be of variable size, from sentences to complete documents and, in general, the vectors are trained to predict words or terms in a paragraph and thus assign a semantic representation.

Then, the method performs a mapping based on probabilities, so that words that have the same meaning are distributed in the same vector space, making it possible to make the semantic distinction between the words in a paragraph. Sequentially, the method maps the paragraphs to different vectors of words, concatenating the vector of the paragraph with several vectors of words present in the paragraph, in order to predict the next word in the context considered. In this way, the variable length of sentences, word order and semantics are taken into account. Both the word and paragraph vectors are trained by descending the stochastic gradient and post-propagation (RUMELHART; HINTON; WILLIAMS, 1986).

It is important to note that while paragraph vectors are unique among paragraphs, word vectors are shared (the vector of a word is the same for all paragraphs that have that word). At the time of prediction, the paragraph vectors are inferred by correcting the word vectors and training the new paragraph vector until convergence. The method can use different strategies for generating paragraph vectors, the main ones being (LE; MIKOLOV, 2014):

a) Distributed Memory Model of Paragraph Vectors (PV-DM): each paragraph is mapped to a unique vector, represented by a column in an array. Each word is also mapped to a unique vector, represented by a column in an array. The concatenation or mean of the paragraph vector with the word vectors is used to predict the next word in a context. The paragraph vector can be considered a pseudo-word and represents the information that is missing in the current context, acting as a memory of the topic of the paragraph in question.

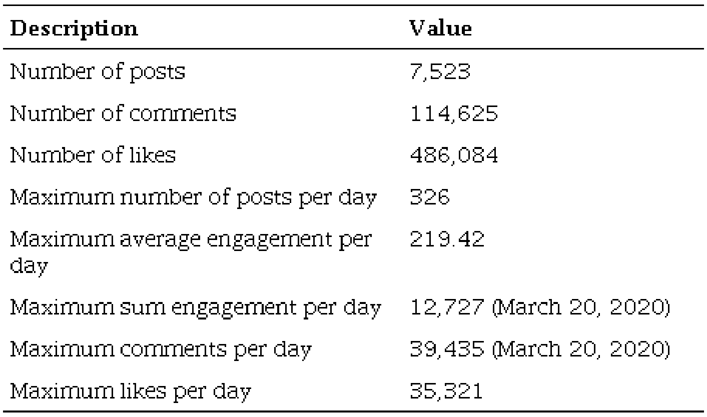
b) Distributed Bag of Words version of Paragraph Vector (PV-DBOW): context words are ignored in the input and are predicted randomly from the paragraph vector.

Sequentially, the implemented model deals with the context. In the vector space in which the documents are mapped, the proximity between vectors represents similar usage patterns, so that words used in the same context are close to each other. This representation considers, therefore, the variable size of the document, the word order and the semantics. The interpretation, then, depends on the set of terms and not on the specific elements of the descriptive text. From the trained and validated model, it is possible to perform the semantic similarity analysis of the posts, which can be visualized through a weighted graph based on the similarity matrix, illustrating the relationships between the posts and the possible detected groups. Through the segmentation of the classes found, a similarity graph is generated between the terms present in the textual corpus, based on the method presented in Bouriche (2005), considering the semantic and syntactic relations. The results are analyzed in view of the epidemiological advances of COVID-19, considering the world reality, as well as the Brazilian scenario of the spread of the virus.

4 Results

With the process of extracting data from the group selected on the social network Facebook, 7523 publications were obtained. Quantitative details about the collected publications are presented in Table 1. After carrying out the cleaning and preprocessing step, the sample was reduced to 5,118 publications, with a total number of 64,854 comments and 301,611 likes. We apply data mining to the data set, for descriptive analysis, pattern extraction and content analysis.

TABLE 1
Quantitative of collected publications from the Facebook group



Description	Value
Number of posts	7,523
Number of comments	114,625
Number of likes	486,084
Maximum number of posts per day	326
Maximum average engagement per day	219.42
Maximum sum engagement per day	12,727 (March 20, 2020)
Maximum comments per day	39,435 (March 20, 2020)
Maximum likes per day	35,321

Source: research data

The results obtained through descriptive analysis are of fundamental importance for the knowledge of the database's characteristics, as well as the detection of specific patterns and behaviors, which, in this case, may be associated with events directly linked to the COVID-19 pandemic, or derived events, such as those linked to socio-economic issues.

During the process of descriptive analysis, the data were normalized to allow comparison with the temporal curve of the number of infected COVID-19 in Brazil. This normalization avoids bias in the analysis for variables with a higher order of magnitude. The normalization was done by dividing the daily value corresponding to the variable of interest xi by the maximum value of this variable in the entire historical series considered x, resulting in a maximum value equal to 1:

$$x_i' = \frac{x_i - min(x)}{max(x) - min(x)}$$

[1]

The first analysis was generated in relation to the evolution of the daily number of posts published in the selected group (Figure 1). It is possible to observe that although there is a relative growth in the number of posts, it does not follow the curve of the number of contaminations in Brazil, with a drop even after the daily peak of posts. However, it is interesting to note some important phenomenological features. The selected group refers to the largest Brazilian public group on the subject on Facebook, however the number of posts remained stable and at a minimum level, less than forty posts between the date of creation of the group and the occurrence of the first case in Brazil, which caused the growth curve to reach one of its peaks, with a total of 263 posts in a single day.

The peak values of daily postings present in Figure 1 demonstrate a clear reaction to events directly linked to the Brazilian scenario, such as the confirmation of the first case of COVID-19 in Brazil, the confirmation of contamination by a member of the presidential executive Brazillian team, and confirmation of the first death due to COVID-19 in Brazil. Such data demonstrate a greater reactive social force in terms of sharing, discussing and socializing opinions and information that can more effectively affect the daily lives of the participants in this social group. In contrast, Figure 2 shows the evolutionary curve in the average daily engagement observed in relation to the growth curve of contamination by COVID-19 in Brazil. The results demonstrate that the growth rate of the average daily engagement accompanies the growth in the number of contaminations. A single point of greater prominence occurred on March 20, 2020, relative to the date of approval of the public state of calamity in Brazil by the Federal Senate (BRASIL, 2020). The number of effectively active participants, that is, those who at some point published at least one post, is 2108 users, which corresponds to 4.5% of the total number of participants in the group, which reveals that there is a considerable participation by other users in terms of monitoring the publications posted in the group.

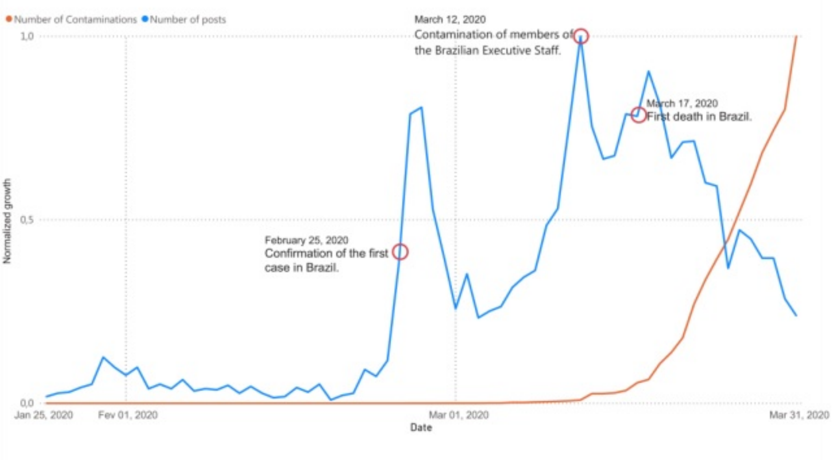


FIGURE 1
Evolution of the daily number of posts
Source: authors

In order to analyze the content published in the posts, an unsupervised method of calculating textual similarity was used, Doc2Vec (LE; MIKOLOV, 2014). From the preprocessed textual database, the method generated a matrix of similarities between the posts. The parameterization of the model, for training purposes on the database called here train-corpus, was carried out based on related works (LEE; WELSH, 2005; MIKOLOV et al., 2013; LE; MIKOLOV, 2014).

With the purpose of validating the model, from train-corpus, vectors of terms were generated for the documents through inference. In this sense, the inference algorithm predicts the terms based on the word vectors, and these new vectors can be compared with the vectors of the trained model. Basically, in this approach, train-corpus is treated as an unknown data by the model and, once similarity between the vectors (inferred and modeled) is identified, a notion of the model's consistency is obtained. Although it is not a real precision value, it is a way of validating how representative the model is for the characteristics of the database documents.

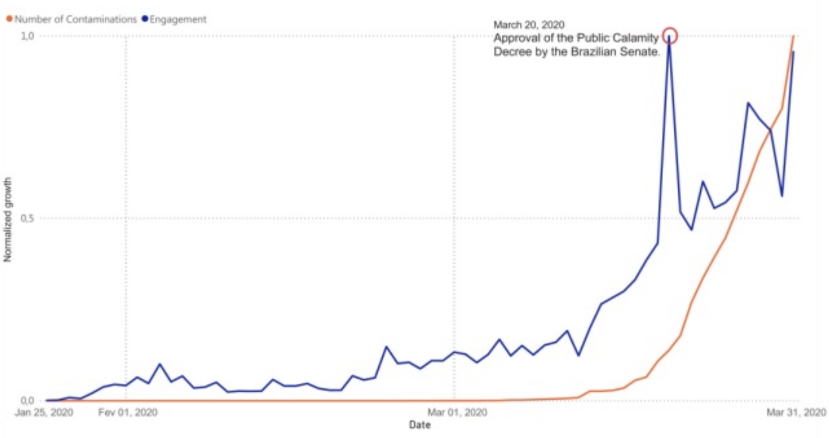


FIGURE 2
Evolution of the daily engagement.
Source: authors

he similarity matrix shows the corresponding values between all elements, that is, considering the existence of n textual elements, n^2 similarity distances are provided. Through this result, a similarity graph was generated, showing the distribution between the posts (Figure 3). In the graph, each node corresponds to a post, an

The similarity matrix shows the corresponding values between all elements, that is, considering the existence of n textual elements, n^2 similarity distances are provided. Through this result, a similarity graph was generated, showing the distribution between the posts (Figure 3). In the graph, each node corresponds to a post, and the edges correspond to the similarity distance between them, that is, the closer the two nodes are, the more similar they are. In this sense, for visualization purposes a threshold was applied to the similarity values, being considered only those greater than 0.7. The application of this threshold is necessary to allow the identification of possible classes of similar elements. The value of 0.7 was defined based on empirical checks. Nodes that are far apart have a lower similarity value compared to the others, and nodes that are between the two classes have similar similarity values between elements that make up each of the classes.

From the identification of the elements that make up each of the groups identified through Doc2Vec, that is, the posts most similar to each other, the most frequent terms within these groups were calculated. Table 2 presents the twenty most frequent terms for both classes. It is possible to observe that the terms related to Class 1 refer to the terms most present in governmental issues and factors directly associated with the pandemic, such as quarantine, hospital, epicenter and death. On the other hand, Class 2 presents terms that are more related to people's daily lives, with emphasis on terms related to food, health and personal care

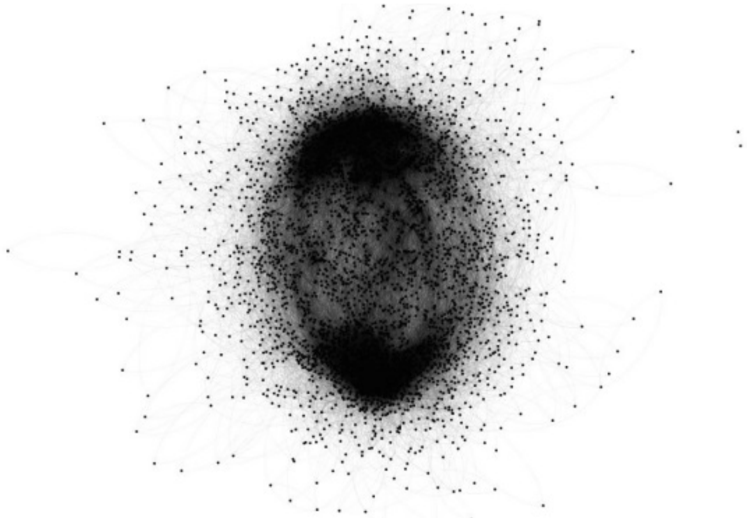


FIGURE 3
Graph of similarity between publications (Doc2Vec)
Source: authors

For a more careful analysis of the information circulating in the group, the similarity graphs were generated. Figure 4 presents the results obtained for Class 1, where the central terms are: brazil, hospital, china, government, and quarantine. Associated with these, there are terms that confirm the characteristic of content related to politics and informative news. Figure 5 shows the results obtained for Class 2. It is interesting to note that in these results the most relevant term is person, which is related to others that refer to family, food, health and personal care.

TABLE 2
Top 20 most frequent words

Class	Top words
Class 1	case, china, brazil, patient, death, hospital, authority, government, test, minister, europe, world, quarantine, agreement, epidemic, president, germany, cruise, epicenter, group.
Class 2	clothes, play, place, body, street, food, meal, hygiene, environment, asthma, husband, person, mask, sleep, community, runny nose, anxiety, will, bar, care.

Source: research data

The results obtained through the similarity graphs demonstrate that in the group analyzed in the selected social network, there are two main patterns of content in accordance with COVID-19, those of an informative and general nature, focused on issues at the level of social groups, and those of a more personal character, focused on the individual's concerns, opinions and desires. Table 3 presents the theme of posts with greater engagement in each of the identified classes, confirming the inferred characteristics.

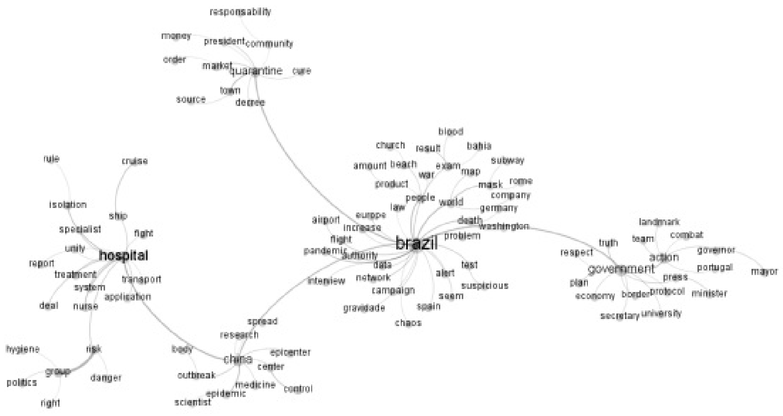


FIGURE 4
Similarity graph for Class 1
Source: authors

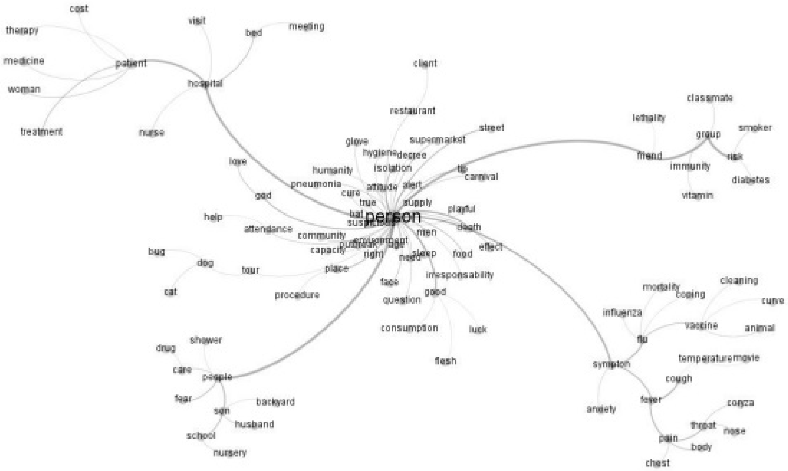
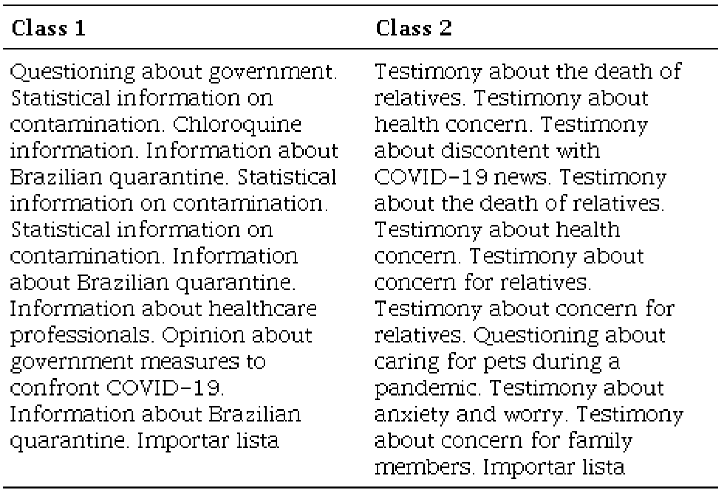


FIGURE 5
Similarity graph for Class 2
Source: authors

TABLE 3
Themes of the ten posts with bigger engagement



Class 1	Class 2
Questioning about government. Statistical information on contamination. Chloroquine information. Information about Brazilian quarantine. Statistical information on contamination. Statistical information on contamination. Information about Brazilian quarantine. Information about healthcare professionals. Opinion about government measures to confront COVID-19. Information about Brazilian quarantine. Importar lista	Testimony about the death of relatives. Testimony about health concern. Testimony about discontent with COVID-19 news. Testimony about the death of relatives. Testimony about health concern. Testimony about concern for relatives. Testimony about concern for relatives. Questioning about caring for pets during a pandemic. Testimony about anxiety and worry. Testimony about concern for family members. Importar lista

Source: research data

5 Conclusion

The article explores the potential for knowledge extraction from the analysis of social interaction in a Facebook group for disseminating information and discussions about the COVID-19 pandemic in Brazil, from January to March 2020. The quantitative analysis of the number of publications and engagement in the explored social group revealed patterns that can be used to represent the reaction of society, as a participant in the information cycle, in view of the evolution of the pandemic in Brazil. In specific periods, there was a strong association between the growth in the number of publications in the group, such as the confirmation of the first death by COVID-19 in Brazil and the contamination of government members, as well as greater engagement in the group from the declaration of state of public calamity.

The qualitative analysis of the publications established the semantic meaning of the information circulating in the group, being an indication of society's understanding of the context of the pandemic. The publications were classified automatically by calculating the similarity of the content using the Doc2Vec method, and the results obtained allowed to map two patterns of information flow in the group: 1) publications that report general interests, represented by informative content about the pandemic and discussions about the coping measures taken by the Brazilian government, as well as comparison of scenarios and 2) publications that report personal perceptions, represented by reports from participants about the impacts of the pandemic on their daily lives, expressing concerns, difficulties in adapting to quarantine, daily routine and new demands.

In general, the results of this study suggest characterizing the group as a means of social convergence, as the circulating information provides its participants with knowledge about the evolution of the pandemic in Brazil and situational awareness. It is also possible that the participants are motivated by the positive psychological effect of the interactionist process, which favors empathy and freedom of expression, by sharing individual feelings and experiences. Based on this motivation, common interests and perceptions are identified and, thus, the network plays an important role in the generation of knowledge and socialization in crisis scenarios.

In this sense, it is understood that the study is significantly relevant, presenting a representative picture of society's response to the evolution of the COVID-19 pandemic in Brazil. In addition, the analyzes can be replicated in other contexts and for different purposes, for example in order to identify the emotions and polarity of feeling about the most engaged publications or even to identify the spread of rumors in the information circulating on a social network.

Acknowledgments

References

AHSAN, M.; KUMARI, M.; SHARMA, T. Rumors detection, verification and controlling mechanisms in online social networks: A survey. **Online Social Networks and Media**, [s.l.], v. 14, p. 100050, 2019.

ARNABOLDI, V.; CONTI, M.; PASSARELLA, A.; DUNBAR, R.I. Online social networks and information diffusion: The role of ego networks. **Online Social Networks and Media**, [s.l.], v. 1, p. 44 – 55, 2017.

BELK, R. Sharing Versus Pseudo-Sharing in Web 2.0. **The Anthropologist**, [s.l.], v. 18, n. 1, p. 7-23, 2014.

BOURICHE, B. L'analyse de similitude. In: ABRIC, J. (ed.) **Méthodes d' #Etude des Représentations Sociales**. Toulose, France: Eres, 2005. p. 221-252.

BRASIL. Decreto-lei nº 6, de 20 de março de 2020. Reconhece, para os fins do art. 65 da Lei Complementar nº 101, de 4 de maio de 2000, a ocorrência do estado de calamidade pública, nos termos da solicitação do Presidente da República encaminhada por meio da Mensagem nº 93, de 18 de março de 2020. Disponible in: https://legis.senado.leg.br/norma/31993957/publicacao/31994188/. Access in: Jul. 31, 2020.

CERCEL, D.; TRAUSAN-MATU, S. Opinion propagation in online social networks: A survey. In: International Conference on Web Intelligence, Mining and Semantics, 4, Thessaloniki, Greece, 2014. **Proceedings** [...]. New York, NY, USA: Association for Computing Machinery, 2014. p. 1-10.

DRAGOVIĆ, N., VASILJEVIC, D.; STANKOV, U.; VUJICI, M. Go social for your own safety! Review of social networks use on natural disasters – case studies from worldwide. **Open Geosciences**, [s. l.], v. 11, n. 1, p. 352-366, 2019.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. From data mining to knowledge discovery in databases. **AI Magazine**, Palo Alto, CA, USA, v. 17, n. 3, p. 37–54, 1996.

FERRARA, E. Disinformation and social bot operations in the run up to the 2017 French presidential election. **First Monday**, Chicago, IL, USA, v. 22, n. 8, 2017.

FISHER, E.; MEHOZAY, Y. How algorithms see their audience: media epistemes and the changing conception of the individual. **Media, Culture & Society**, [s. l.], v. 41, n. 8, p. 1176–1191, 2019.

GREENELTCH, N. **Python Data Mining Quick Start Guide**. Birmingham, UK: Packt Publishing, 2019.

GRUEBNER, O.; LOWE, S.R.; SYKORA, M.; SHANKARDASS, K.; SUBRAMANIAN, S.; GALEA, S. Spatio-temporal distribution of negative emotions in New York city after a natural disaster as seen in social media. **International Journal of Environmental Research and Public Health**, Basel, Switzerland, v. 15, p. 1-12, 2018.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: concepts and techniques. Burlington, Massachusetts, USA: Morgan Kaufmann Publishers, 2016.

HUSSAIN, M.N.; TOKDEMIR, S.; AGARWAL, N.; AL-KHATEEB, S. Analyzing disinformation and crowd manipulation tactics on Youtube. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018. **Proceedings** [...]. Piscataway, NJ, USA: IEEE, 2018. p. 1092-1095.

JARMUL, K.; LAWSON, R. **Python Web Scraping**. Birmingham, UK: Packt Publishing, 2017.

JUST, N.; LATZER, M. Governance by algorithms: reality construction by algorithmic selection on the Internet. **Media, Culture & Society**, [s. l.], v. 39, n. 2, p. 238-258, 2017.

KANE, G.C.; ALAVI, M.; LABIANCA, G.; BORGATTI, S.P. What's different about social media networks? A framework and research agenda. **MIS Quarterly**, Minnesota, USA, v. 38, n. 1, p. 274-304, 2014.

KEIM, M; NOJI, E. Emergent use of social media: A new age of opportunity for disaster resilience. **American Journal of Disaster Medicine**, Weston, MA, USA, v. 6, n. 1, p. 47-54, 2011.

KEMP, S. **Digital 2020**: Global digital yearbook. [s. l.]: Kepios Pte. Ltd., 2020. Disponible in: https://datareportal.com/reports/digital-2020-global-digital-yearbook. Access in: Jul. 31, 2020.

KIM, J.; BAE, J.; HASTAK, M. Emergency information diffusion on online social media during storm Cindy in U.S. **International Journal of Information Management**, [s. l.], v. 40, p. 153-165, 2018.

KRYVASHEYEU, Y; CHEN, H; OBRADOVICH, N; MORO, E; VAN HENTENRYCK, P; FOWLER, J.; CEBRIAN, M. Rapid assessment of disaster damage using social media activity. **Science Advances**, Washington, DC, USA, v. 2, n. 3, p. 1-11, 2016.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: International Conference on Machine Learning, 31, Beijing, China, 2014. **Proceedings** [...]. v. 32. Beijing, China, 2014, p. 1-9.

LEE, D.; WELSH, M. An empirical evaluation of models of text document similarity. In: Annual Conference of the Cognitive Science Society, 27, Stresa, Italy, 2005. **Proceedings** [...]. Mahwah, NJ, USA: Lawrence Erlbaum Associates Inc., 2005. p. 1254-1259.

LEVY, P. Virtual communities and information services: an overview. **VINE**, [s. l.], v. 7, n. 5, p. 3-9, 1997.

LI, N.; DAS, S.K. Efficiently discovering users connectivity with local information in online social networks. **Online Social Networks and Media**, [s. l.], v. 16, p. 100062, 2020.

LI, X.; XIE, Q.; JIANG, J.; ZHOU, Y.; HUANG, L. Identifying and monitoring the development trends of emerging technologies using patent analysis and twitter data mining: the case of perovskite solar cell technology. **Technological Forecasting and Social Change**, [s. l.], v. 146, p. 687-705, 2019.

LIPSCHULTZ, J.H. **Free expression in the age of the Internet:** social and legal boundaries. Abingdon, UK: Routledge, 2018.

LIU, T.; ZHANG, H.; ZHANG, H. The impact of social media on risk communication of disasters - a comparative study based on sina weibo blogs related to tianjin explosion and typhoon pigeon. **International Journal of Environmental Research and Public Health**, Basel, v. 17, n. 3, p. 1-17, 2020.

LIU, Y.; GAYLE, A.; WILDER-SMITH, A.; ROCKLÖV, J. The reproductive number of covid-19 is higher compared to SARS coronavirus. **Journal of Travel Medicine**, v. 27, n. 2, p. 1-4, 2020.

LOGAN, R.K. **Understanding new media**: Extending Marshall McLuhan. Second edition. Bern, Switzerland: Peter Lang Publishing, 2016.

MELO, C.M.L.; SILVA, G.A.S.; MELO, A.R.S.; DE FREITAS, A.C. COVID-19 pandemic outbreak: the Brazilian reality from the first case to the collapse of health services. **Annals of the Brazilian Academy of Sciences**, Rio de Janeiro, Brazil, v. 94, n. 4, p. 1-14, 2020.

MCLUHAN, M. **Understanding Media**: The Extensions of Man. Cambridge, MA, USA: MIT Press, 1994.

MIKOLOV, T.; SUTSKEVER, I.; CHEN K.; CORRADO G.; DEAN J. Distributed representations of words and phrases and their compositionality. In: International Conference on Neural Information Processing Systems, 26, Stateline, Nevada, USA, 2013. **Proceedings** [...]. Red Hook, NY, USA: Curran Associates Inc., 2013. p. 3111-3119.

O'REILLY, T. **What Is Web 2.0** - Design Patterns and Business Models for the Next Generation of Software. Newton, Massachusetts, USA: O'Reilly Publishing, 2005.

PAN F.; YE, T.; SUN, P.; GUI, S.; LIANG, B.; LI, L.; ZHENG, D.; WANG, J.; HESKETH, R.L., YANG, L.; ZHENG, C. Time course of lung changes on Chest CT during recovery from 2019 novel coronavirus (Covid-19). **Radiology**, Oak Brook, IL, USA, v. 295, n. 3, p. 715-721, 2020.

PENNI, J. The future of online social networks (osn): A measurement analysis using social media tools and application. **Telematics and Informatics**, [s. l.], v. 34, n. 5, p. 498-517, 2017.

PROVOST F.; FAWCETT, T. **Data Science for Business**: What You Need to Know about Data Mining and Data-Analytic Thinking. Newton, Massachusetts, USA: O'Reilly Media, 2013.

RIVEST, R. **RFC1321**: The MD5 Message-Digest Algorithm. Marina del Rey, CA, USA: RFC Editor, 1992.

SALLOUM, S.; AL-EMRAN, M.; SHAALAN, K. Mining social media text: extracting knowledge from Facebook. **International Journal of Computing and Digital Systems**, Bahrain, v. 6, n. 2, p. 73-81, 2017.

SREEJESH, S.; PAUL, J.; STRONG, C.; PIUS, J. Consumer response towards social media advertising: effect of media interactivity, its conditions and the underlying mechanism. **International Journal of Information Management**, [s. l.], v. 54, p. 102-155, 2020.

STATISTA. Increased media device usage due to the coronavirus outbreak among internet users worldwide as of march 2020, by country. *In*: STATISTA. [Hamburg: Statista GmbH], 2020a. Disponible in: https://www.statista.com/statistics/1106607/device-usage-coronavirus-worldwide-by-country/. Access in: Jul. 31, 2020.

STATISTA. Leading countries based on number of Facebook users as of January 2020. *In*: STATISTA. [Hamburg: Statista GmbH], 2020b. Disponible in: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/. Access in: Jul. 31, 2020.

TAN, P.; STEINBACH, M.; KARPATNE, A.; KUMAR, V. **Introduction to Data Mining**. Second edition. London, UK: Pearson, 2005.

TANG, J.; SUN, J.; WANG, C.; YANG, Z. Social influence analysis in large-scale networks. In: ACM SIGKDD international conference on Knowledge discovery and data mining, 15, 2009, Paris, France. **Proceedings** [...]. Paris, France: ACM, 2009. p. 807-816.

WANG H.; CASTANON, J.A. Sentiment expression via emoticons on social media. In: IEEE International Conference on Big Data, Santa Clara, CA, USA, 2015. **Proceedings**[...]. Piscataway, NJ, USA: IEEE, 2015. p. 2404-2408.

WORLD HEALTH ORGANIZATION. **Coronavirus disease 2019 (covid-19) situation report – 69**. Genève, Switzerland: World Health Organization, 2020a.

WORLD HEALTH ORGANIZATION. **Who Director-General's opening remarks at the media briefing on Covid-19 – 11 March 2020**. 11 mar. 2020b. Disponible in: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020?fbclid=IwAR1kA7MQ8c5t-th5B6VoZWiaPDNP6X8QHEK-9ICjXPd5tNcvPU3fIH34MT4/. Access in: Jul. 31, 2020.

WU, Z.; MCGOOGAN J.M. Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China. **JAMA**, Chicago, IL, USA, v. 323, n. 1, p. 1239-1242, 2020.

RUMELHART, D.; HINTON G.; WILLIAMS, R. Learning representations by back-propagating errors. **Nature**, [s. l.], v. 323, p. 533-536, 1986.

Glossary

: