EPiC Series in Computing
Volume 82, 2022, Pages 112–121
Proceedings of 37th International Conference on Computers and Their Applications
EPiC Computing

A Topic Modeling Method for Analysis of Short-Text Data in Social Media Networks

Ian Santos, Luciana Rech, Ricardo Moraes

Universidade Federal de Santa Catarina, Florianópolis, Brasil
ian.macedo@posgrad.ufsc.br, luciana.rech@ufsc.br,
ricardo.moraes@ufsc.br

Abstract

Currently, many short texts are published online, especially on social media platforms. High impact events, for example, are highly commented on by users. Understanding the subjects and patterns hidden in online discussions is a very important task for contexts such as elections, natural disasters or major sporting events. However, many works of this nature use techniques that, despite showing satisfactory results, are not the most suitable when it comes to the short texts on social media and may suffer a loss in their results. Therefore, this paper presents a text mining method for messages published on social media, with a data pre-processing step and topic modeling for short texts. For this paper, we created a data set from real world tweets related to COVID-19 that is openly available[1] for research purposes.

1 Introduction

On March 11, 2020, the World Health Organization (WHO) officially declared that the outbreak of a new coronavirus that was spreading rapidly across several countries had reached pandemic level [1]. In order to prevent the spread of the virus, the need for people to isolate themselves arose. Consequently, many have turned to social media in order to, among other things, keep in touch with others, understand what is happening, and discuss how to adapt to this new reality.

The intense use of social media, such as Twitter or Facebook, accelerates the process of exchanging information and expressing opinions about events and health crises [2]. As a consequence, the COVID-19 pandemic has been an intensely discussed topic since January 2020 to date.

Textual data collected from Twitter is valuable for showing public conversations related to various topics, as well as real-time updates and news during global pandemics [3]. It can even be a source for studies of infodemiology, that is, the "pandemic" of information, be it facts, rumors or fake news. Since the beginning of the COVID-19 pandemic, there has been an effort by the scientific community to study, analyze and understand how people were reacting to this whole situation [4, 5, 6]. The ordinary citizens,

[1] https://zenodo.org/record/5781643#.YbkYw73MJPY

news portals, high-ranking politicians and even government organizations somehow used Twitter as a communication channel, where they constantly shared COVID-19-related news to the public.

Many studies of this nature use text mining techniques such as topic modeling to analyze the hidden patterns between the words of a corpus of documents, that is, a collection of texts [7]. However, most of these works use techniques such as Latent Dirichlet Allocation (LDA), which is not suitable for texts from social media. This is because most texts on social media are short and have a limited context, suffering from the sparsity of data.

Therefore, it is necessary to use adequate techniques according to these characteristics. The Biterm Topic Modeling (BTM) is an alternative to this problem, as it performs the modeling considering the whole context of word pair co-occurrence in the corpus, different from traditional methods that analyze the co-occurrence of words at the document level. This means that in documents with short texts, there is a limitation in the context in which the modeling takes place.

The objective of this work is to show a method to examine discussions related to COVID-19 and the interests of users in social media, based on messages posted by them. For this, we prepared two corpuses of texts collected from Twitter, related to COVID-19 and from different periods. Then, we performed a BTM topic modeling technique for short texts on these sets of texts to highlight the issues discussed by the users.

Thus, this research demonstrates a way to investigate the textual content being generated by users, by performing an analysis on the data generated by social media users. In the context of this research, we did not consider the reach of the tweets, that is, how much engagement a publication had or the number of times a particular tweet was shared. We only considered the content of the texts.

2 Background

The method presented in this paper consists in a combination of techniques to be applied to a collection of texts published by Twitter users to identify latent topics in online discussions. These techniques involve natural language processing and text mining, unsupervised machine learning, and the Renyi entropy calculation routine.

Normally, documents are texts composed of several topics, with a topic being a set that represents the probability distribution of words with a semantic relationship to each other [8]. Topic modeling is a text mining tool that encompasses statistical techniques that can efficiently reveal latent issues that are being addressed in a text corpus [9]. Unsupervised machine learning methods such as LDA are widely used by researchers who analyze data from unstructured texts such as news articles or social media publications.

Given a corpus of documents, where each document contains a text, traditional modeling such as LDA identifies word co-occurrence patterns at the document level to reveal topics [8]. However, when it comes to short texts like those that are mostly present in social media, this type of solution suffers from sparse data, that is, the co-occurrence of words becomes less common in each document due to the small size of the texts [10]. Among other adversities, the small amount of text makes it difficult for topic modeling to identify the meaning of ambiguous words in each context.

Thus, the topics identified in the modeling are more repetitive and include a lot of noise that does not contribute to the semantic value of the texts, especially when there is no pre-processing of the texts. Several researches and studies have already proposed solutions to this problem, which include new algorithms and data pre-processing procedures that show more effective results when it comes to modeling topics in short texts.

2.1 Biterm Topic Modelling (BTM)

Short texts are widely found in social media messages, especially on Twitter, where there is a character limit for each tweet. BTM is a modeling technique adapted for modeling topics in short texts, using the co-occurrence relationship of words to efficiently solve the data sparsity problem. BTM learns topics by modeling word-word co-occurrence patterns called biterms [11].

The main idea of BTM is to learn topics in a collection of short texts based on the aggregated biterms of the entire corpus, in order to solve sparsity problems of a single document. In [8], it is described that the entire corpus is a mixture of topics, where each biterm is extracted from a specific topic independently. Thus, the probability of a biterm extracted from a specific topic is later captured by the chances that both words in the biterm are extracted from the topic. In this way, all biterms of the entire corpus are brought together to learn topics, taking full advantage of the various global patterns of word co-occurrence to better reveal latent topics.

An important parameter in any topic modeling is the number of topics to be identified from a corpus. This is an integer value that depends on factors such as corpus size. To determine the optimal number of topics for a modeling, the Renyi entropy calculation was adopted in this work.

2.2 Renyi entropy calculation routine

As textual data becomes more abundant over the Internet, new methods and tools such as topic modeling emerge to process and model this data [12]. However, this is a tool in constant evolution, which makes it difficult to choose the ideal parameters for such models.

Entropy is a thermodynamic concept to represent the disorder of a system, that is, the greater the alteration of the state of a system, the greater its disorder [13]. The authors also declared that entropies can be used to quantify the diversity, uncertainty, and randomness of a system, as it can be used in ecology and statistics, for example, as an important index of diversity in a data set. According to the principle of maximum entropy [14], entropy is considered as negative information, therefore, the maximum entropy corresponds to the minimum of information. It is assumed that the ideal number of topics corresponds to the maximum information received as a result of topic modeling.

In [15], it was studied the behavior of topic modeling with changes in the number of topics based on the thermodynamic concepts of entropy, and it was presented an application of Renyi entropy for the optimization of topic modeling, based on the maximum information approach and on the principle of T invariance. According to the author, it is possible to investigate the distribution behavior of a set of words when the number of topics changes.

Then, this paper presents an approach formulated according to the following provisions [15, 16]: (1) the number of words and documents is a constant, that is, there is no change in volume. (2) a topic is a state that each word and document can assume. Words and documents can belong to different topics with different probabilities. (3) the thermodynamic information system is open and exchanges energy with the external environment when changing temperature, which is understood as the T number of topics. This value is the parameter to be determined when looking for the minimum non-extensive entropy of the system. (4) As a measure of the non-equilibrium of such an information system, one can use the entropy difference $\Lambda s = S - S_0$ [17], where S_0 is the entropy of the state, and S is the entropy of the non-equilibrium state. As an analogy, a function can be formulated based on the difference in free energies: $\Lambda F = F(T) - F_0$, where F_0 is the free energy of the initial state (chaos), and F(T) the free energy in a given value of T. (5) Since topic modeling is a method for finding the distribution of words, the number of distributions is a variable parameter. The ideal number of these distributions corresponds to the situation where the minimum entropy is reached. (6) it is assumed that the equilibrium state of an information system is due to the fact that a set of words with a high probability stops changing with a change in the number of topics.

2.3 Related work

Analysis of textual content on Twitter was performed in several studies, especially during the COVID-19 pandemic. The studies cover approaches such as manual [5] or descriptive [6] analysis of Twitter content to interpret the data. Other studies use unsupervised machine learning techniques, such as [4] that present a study on the relationship of topic modeling and sentiment analysis. For that, they divided a data set of tweets in the context of COVID-19 into three parts, covering positive, neutral and negative sentiments. Then, the Non-Negative Matrix Factorization (NMF) algorithm generates a list of topics for each set of sentiments, in order to perform a topic analysis on each set. Finally, they found that topic modeling together with sentiment analysis complement each other for a better understanding of opinions and topics discussed online.

However, the work of [4], as well as most works of this nature, use traditional algorithms such as LDA [7], which can limit the modeling potential. It is also very common for analysis like those to be carried out with data in English, highlighting a shortage of studies in other languages.

Another work with these characteristics was published in [3], which sought to identify the main topics related to the COVID-19 pandemic discussed by Twitter users. Using tools like Python and the PostgreSQL database, together with a set of predefined terms (such as "corona" and "COVID-19"), they extracted text and metadata from English tweets from February 2, 2020, to March 15, 2020. LDA was used to model topics. They performed a sentiment analysis, and calculated the interaction rate per topic. At the end, 12 topics were identified, where, on average, 10 topics had positive feelings, and 2 topics with negative feelings (deaths caused by COVID-19 and increased racism). The topic with the highest average of likes was 15.4 (economic losses) and the lowest average was 3.94 (travel bans).

Based on these facts, this work proposes a method with a data pre-processing step, so that the topic modeling can be performed in documents with short texts and using another solution, different from the one traditionally adopted. To demonstrate the results, we applied the method on two distinct tweet data sets relating to the COVID-19 pandemic.

3 Methodology

We gathered the tweets used in this study from the data set of [18] which, since January 2020, has gathered more than one billion tweets related to COVID-19 in several languages, including Portuguese. For this work, all tweets from March 2020 and March 2021 were selected. The data mining process on Twitter followed the pipeline illustrated in Figure 1.

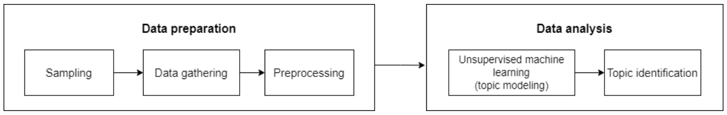


Figure 1: The full pipeline followed to mine Twitter data.

Data preparation takes place in three steps: (1) sampling, (2) data gathering, and (3) raw data preprocessing. Data analysis includes the application of an unsupervised machine learning algorithm and a classification of the results obtained.

When it comes to a large-scale set of texts, a qualitative analysis in itself is a big challenge. Unsupervised learning allows exploratory analysis of texts to be viable for purposes such as social science research. In this work, we applied unsupervised learning to identify topics that stand out in the texts. Unsupervised machine learning is an approach used to look for patterns in the data, and derive probabilities of clusters, based on the data in the text. Unsupervised data learning was used in this work, as it is widely adopted in studies involving unstructured texts where there is little information about them [19].

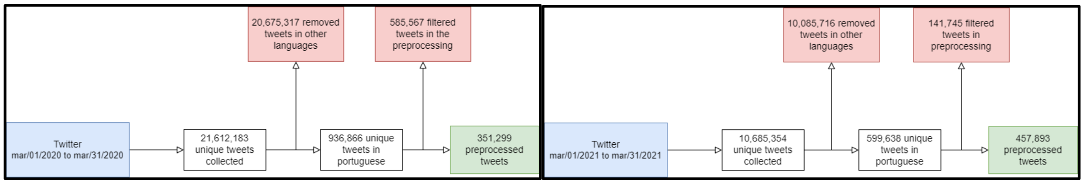
In their work, [18] collect and assembles tweets based on terms and hashtags related to COVID-19, such as #coronavirus, #COVID19, and #corona among many others. As time went by, they incorporated more terms related to the pandemic into the job search process. A social media mining toolkit was used to monitor the Twitter Stream API, based on some COVID-19 related keywords. All tweets with the indicated keywords were collected and processed to a format that could be included in the data set. Other contributors, mainly from non-English languages, contributed to the construction of the data set using similar methods. At the end, all collected tweets are aggregated into a single set.

Then, the texts obtained undergo preprocessing steps. The first part of the process follows the procedure presented by [20]: removal of the hashtag symbol, user mentions, and URLs; removal of characters that are not part of the ASCII code; and removal of special characters, punctuation, and stop words, as these do not contribute to the semantic meaning of the messages. In addition to that, remove any HTML tags, emojis, and other non-alphanumeric symbols, isolated numbers, repeated words in the same document, words that rarely appear in the corpus (less than ten times) and repeated documents, leaving only one copy in the corpus.

As previously mentioned, Renyi entropy was used for modeling the number of topics [15]. Then, topic modeling was performed multiple times on the preprocessed data set, each time modeling a different number of topics. The main idea is to perform the Renyi entropy calculation for each model, and then identify the model that returns the lowest value for the calculation. The minimum value of the Renyi entropy indicates the ideal number of topics.

4 Results

Texts taken from social media have many elements like URLs, emojis, and special characters that are unnecessary for topic modeling, and must be filtered. At first, 21,612,183 unique tweets from March 2020 and 10,685,354 unique tweets from March 2021 were collected, regardless of language. After selecting only the texts in Portuguese, 936,866 tweets from March 2020 and 599,638 tweets from March 2021 were obtained. With the preprocessing of the texts, the samples went to 351,299 documents from 2020 and 457,893 documents from 2021. To maintain the balance between the comparison of the two sets, any modeling in the 2021 corpus will be done with the same number of documents as in the 2020 corpus. Figures 3 and 4 show the flow of preprocessing activities in each data set.


20,675,317 removed tweets in other languages
585,567 filtered tweets in the preprocessing
Twitter mar/01/2020 to mar/31/2020
21,612,183 unique tweets collected
936,866 unique tweets in portuguese
351,299 preprocessed tweets
10,085,716 removed tweets in other languages
141,745 filtered tweets in preprocessing
Twitter mar/01/2021 to mar/31/2021
10,685,354 unique tweets collected
599,638 unique tweets in portuguese
457,893 preprocessed tweets

Figures 3 and 4: March 2020 and March 2021 tweets preprocessing flow, from left to right.

For this work, the text preprocessing was conducted using the Python 3 programming language, using the Pandas and NumPy libraries. We created a script to filter the content in the texts of the data set, seeking to subtract unwanted elements in the text, as explained in section 3 of this work. In addition, we identified the most popular unigrams (word units) and bigrams (word pairs) of the data sets, which are visualized as word clouds in Figures 5 and 6.

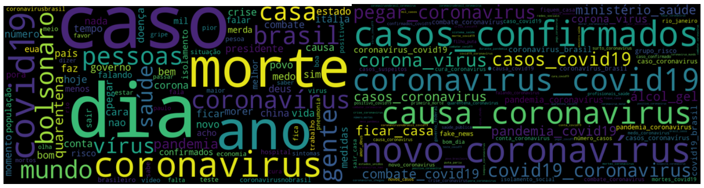
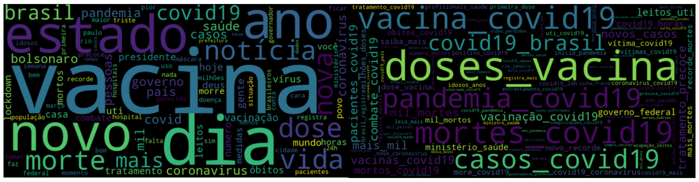


Figure 5: Word clouds of the most popular unigrams and bigrams of 2020, from left to right.



Figures 6: Word clouds of the most popular unigrams and bigrams from 2021, from left to right.

As of March 2020, the most popular unigrams consist of words like: death, case, people, home, crisis, world. The most popular bigrams are: confirmed cases, covid19 cases, covid19 combat, stay home, fake news. For March 2021, the most popular unigrams are: vaccine, state, Brazil, death, thousand, dose, life, news, treatment. Among the bigrams there are terms such as: vaccine doses, covid19 vaccine, covid19 deaths, federal government, a thousand dead, ministry of health. Bigrams are represented with an underscore between them for easier viewing.

Topic modeling gathers pairs of words that appear frequently and organizes them into topics, groups of words that are semantically related to each other. For that, it is necessary to manually set the number of topics. The Renyi entropy calculation routine [15] was adopted to calculate the most appropriate number of topics. For each data set, several distinct models were created covering from two to twenty topics. Thus, it was identified through the Renyi entropy that the most suitable number of topics would be, on average, 10 topics. Tables 1 and 2 present, respectively, the 10 topics identified for March 2020 and March 2021, together with their most popular terms. For better understanding, we have provided an extra column in each table with a translated version of each term.

Table 1: Topics identified from March 2020 and their related terms.

Topic	Topic Related Terms (English)	Topic Related Terms (Portuguese)
Coronavirus Spread	coronavirus, Brazil, people, cases, Bolsonaro, world, home, virus, new	coronavírus, Brasil, pessoas, casos, Bolsonaro, mundo, casa, gente, vírus, novo
Understanding COVID-19	health, measures, pandemic, combat, prevention, information, cause, actions, avoid, quarantine	saúde, medidas, pandemia, combate, prevenção, informações, causa, ações, evitar, quarentena
Pandemic Fatalities	deaths, confirmed, number, Italy, thousand, country, day, state, China, today	mortes, confirmados, número, Itália, mil, país, dia, estado, China, hoje
Economic difficulties	crisis, government, Brazil, economy, money, helping, millions, companies, fund, salary	crise, governo, Brasil, economia, dinheiro, ajudar, milhões, empresas, fundo, salário
Awareness and recommendations	coronavirusinbrasil, stayathome, link, video, Globo, times, live, moment, Instagram, Twitter	coronavirusnobrasil, fiqueemcasa, link, video, Globo, tempos, live, momento, Instagram, Twitter
Prevention of	stay, well, hands, risk, isolation, life,	ficar, bem, mãos, risco, isolamento, vida,

COVID-19	avoid, alcohol, gel, God	evitar, álcool, gel, Deus
Understanding the symptoms	disease, flu, pneumonia, symptoms, corona, risk, vaccine, tests, treatment, study	doença, gripe, pneumonia, sintomas, corona, risco, vacina, testes, tratamento, estudo
Politicization of COVID-19	Bolsonaro, president, government, little flu, press, media, governors, Trump, nothing, irresponsible	Bolsonaro, presidente, governo, gripezinha, imprensa, mídia, governadores, Trump, nada, irresponsável
Advance of COVID-19	positive, hospital, exam, suspicion, negative, result, patient, first victim, state	positivo, hospital, exame, suspeita, negativo, resultado, paciente, primeira, vítima, estado
Anxiety about COVID-19	house, stay, pick up, quarantine, people, fear, shit, mother, go out, life	casa, ficar, pegar, quarentena, pessoas, medo, merda, mãe, sair, vida

Table 2. Topics identified from March 2021 and their related terms.

Topic	Topic Related Terms (English)	Topic Related Terms (Portuguese)
General discussions about COVID-19	covid19, Brazil, deaths, pandemic, vaccine, health, people, cases, thousand, day	covid19, Brasil, mortes, pandemia, vacina, saúde, pessoas, casos, mil, dia
Disapproval of President's Attitudes	vaccine, covid, Bolsonaro, people, president, don't, take, say, son, bitch	vacina, covid, Bolsonaro, povo, presidente, não, tomar, falar, filho, puta
Medicine Discussions	treatment, early, patients, ivermectin, use, chloroquine, doctors, efficacy, kit, azithromycin	tratamento, precoce, pacientes, ivermectina, uso, cloroquina, médicos, eficácia, kit, azitromicina
Afflictions and Fatalities caused by COVID-19	dies, victim, hospitalized, complications, ICU, death, family, mother, Olímpio	morre, vítima, internado, hospital, complicações, UTI, morte, família, mãe, Olímpio
News about COVID-19	deaths, Brazil, cases, a thousand, deaths, day, record, new, record, 24h	mortes, Brasil, casos, mil, óbitos, dia, recorde, novos, registra, 24h
Sorrow about the pandemic in Brazil	Brazil, life, mask, deaths, lives, moment, sad, country, lack, vaccine	Brasil, vida, máscara, mortes, vidas, momento, triste, país, falta, vacina
COVID-19 vaccination	vaccination, vaccines, doses, elderly, first, government, purchase, immunization, Butantan, Anvisa	vacinação, vacinas, doses, idosos, primeira, governo, compra, imunização, Butantan, Anvisa
Fight against COVID-19	more, beds, pandemic, measures, ICU, patients, government, combat, lockdown, decree	mais, leitos, pandemia, medidas, UTI, pacientes, governo, combate, lockdown, decreto
Politicization of COVID-19	Bolsonaro, government, deaths, people, vaccines, genocide, money, minister, outbolsonaro, Lula	Bolsonaro, governo, mortes, povo, vacinas, genocida, dinheiro, ministro, forabolsonaro, Lula
Coronavirus variant dissemination	coronavirus, country, variant, cases, new, study, Brazilian, WHO, data, transmission	coronavírus, país, variante, casos, nova, estudo, brasileira, OMS, dados, transmissão

Due to the uniqueness of some aspects in the Portuguese grammar, some translations had to be adapted for a better understanding. With the results of the topic modeling processes, it was obtained, for each topic, a large collection of the most relevant words for each topic, where the elements of each set of terms have some form of semantic relationship with each other. Thus, based on the words found in each topic, it was possible to identify the subject covered by each one.

Some words were relevant in several topics, so they can be seen in more than one topic. For the visualization in Tables 1 and 2, we sought to prioritize the visualization of words that had not yet appeared in other topics, so that more relevant terms could be shown.

5 Discussion

The steps taken to carry out this work, from data collection to the execution of the topic modeling process, were crucial to obtain the results shown in section 4. To demonstrate the capacity of the method adopted, the process was carried out in two sets of data related to the same theme, but from different periods. In such a manner, it is possible to see the differences in the discussions of each period.

This study has prompted public discussions related to COVID-19. Online users were analyzed from messages in Portuguese published on the Twitter microblogging platform. In March 2020, Twitter users expressed their concerns about the ways to deal with COVID-19 pandemic and, while March 2021 has a more politicized scenario, in addition to several topics presenting more pessimistic scenarios with terms like "death" and even about a new variant of the virus.

The results show an interesting behavior regarding online discussions. In the preprocessing stage of data sets, there was a significant decrease in the number of records in both data sets. The March 2020 pool originally had 936,866 tweets in Portuguese, of which 585,567 were filtered. The period of March 2021 showed a similar behavior, with 599,638 tweets in Portuguese at first, of which 141,745 were filtered. We observed that the great decrease in the amount of tweets happened, mainly, when the duplicate records of each set were eliminated.

The month of March 2020, was the period where the WHO declared the outbreak of the new coronavirus as a pandemic [1]. Hence, there was a great effort by social media users to discuss and understand what it meant, how to protect themselves and how it would affect everyone.

A year later, in March 2021, the discussions were related to issues arising from the pandemic. Drugs such as ivermectin and chloroquine, as well as their effectiveness in the treatment of COVID-19, were debated. Disapproval of the president of Brazil and news reporting thousands of deaths a day are also topics of discussion. Furthermore, there were lamentations about the situation in Brazil, even though vaccination was also a hot topic.

It is important to highlight the characteristics and limitations found in the work. First, the texts used in this research are derived from a data set of tweets related to COVID-19, and only one source was used for data collection. Despite being a massive data set, the Portuguese language was a significant language of its composition, although some keywords more specific to Portuguese, as well as new terms that eventually arose with the evolution of the discussions were not included or were included later in the capture of real-time tweets. Second, although this work involves every kind of tweets in Portuguese, the results show that, in that language, the focus of the discussions was Brazil, with direct quotes to the country, its political personalities and institutions. Therefore, it is fair to assume that Brazilians are the majority in Portuguese Twitter discussions. Third, the opinions of Twitter users do not represent the entire population of Brazil. The topics and tweets involved in these analyses show the opinions and reactions of Twitter users to COVID-19 in the two periods analyzed. Finally, it is worth to mention the data sets produced and used in this work are openly available to be accessed and used by anyone for research purposes.

6 Conclusion

Data from social media, along with machine learning applications and natural language processing, are great allies in infodemiology studies. This becomes even more evident when considering the context of a global pandemic such as COVID-19. Analysis of public discussions is useful to reveal patterns that are unnoticeable.

The results highlight the most relevant topics and show the interests and concerns of users in the initial period of the pandemic and how they were a year later. The word "death", and its derivations, appear in several topics, especially the most tragic ones. The president of Brazil also appears in some topics, accompanied by insults. As the pandemic situation evolved, more issues arose. In March 2020, for example, there were many topics of a more informative aspect, while March 2021 has more topics of a tragic nature, and also about alleged drugs and even about a variant of the virus.

Extrapolating the pandemic scenario, the method presented in this work can also be applied in other contexts. Other events such as natural disasters, elections, or sporting events can be investigated so that their latent discussions are revealed. Furthermore, since this is a statistical technique to model sets of words with a semantic relationship to each other based on their occurrence in the text, therefore, this method is not limited to English, Portuguese, or any other language.

In future research, we seek to compare the method presented in this paper to other similar methods proposed in the literature. We also intend to analyze the performance of other topic modeling techniques for short texts while using this methodology, besides attesting the performance of this method in other datasets of different languages, such as English and Spanish. Furthermore, we aspire to include a previous step to this method, that covers the text data retrieval process directly from the source.

7 Acknowledgment

For the support and incentives provided during the execution of this work, the authors would like to express their gratitude to the Brazilian research agencies, especially *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* – Brasil (CAPES) and the *Demanda Social* program.

References

[1] WHO, "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," 2020. [Online]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[2] Hans Rosenberg and Shahbaz Syed and Salim Rezaie, "The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic," in Canadian journal of emergency medicine, Cambridge University Press, 2020, pp. 418-421.

[3] Alaa Abd-Alrazaq and Dari Alhuwail and Mowafa Househ and Mounir Hamdi and Zubair Shah, "Top concerns of tweeters during the covid-19 pandemic: infoveillance study," in *Journal of medical Internet research*, JMIR, 2020.

[4] Matheus Adler Soares Pinto and Antonio Fernando Lavareda Jacob Junior and Antonio José G Busson and Sérgio Colcher, "Relacionando Modelagem de Tópicos e Classificação de Sentimentos para Análise de Mensagens do Twitter Durante a Pandemia da COVID-19," in Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web, SBC, 2020, pp. 61-64.

[5] Ramez Kouzy and Joseph Abi Jaoude and Afif Kraitem and Molly B El Alam and Basil Karam and Elio Adib and Jabra Zarka and Cindy Traboulsi and Elie W Akl and Khalil Baddour, "Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter," in Cureus, Cureus Inc., 2020.

[6] Lisa Singh and Shweta Bansal and Leticia Bode and Ceren Budak and Guangqing Chi and Kornraphop Kawintiranon and Colton Padden and Rebecca Vanarsdall and Emily Vraga and Yanchen Wang, "A first look at COVID-19 information and misinformation sharing on Twitter," in arXiv preprint arXiv:2003.13907, 2020.

[7] Oguzhan Gencoglu, "Large-scale, language-agnostic discourse classification of tweets during COVID-19," in Machine Learning and Knowledge Extraction, Multidisciplinary Digital Publishing Institute, 2020, pp. 603-616.

[8] Xiaohui Yan and Jiafeng Guo and Yanyan Lan and Xueqi Cheng, "A biterm topic model for short texts," in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 1445-1456.

[9] Robertus Nugroho, "A survey of recent methods on deriving topics from Twitter: algorithm to evaluation," in Knowledge and Information Systems, Springer, 2020, pp. 2485-2519.

[10] Xiaobao Wu and Chunping Li, "Short text topic modeling with flexible word patterns," in 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1-7.

[11] Xueqi Cheng and Xiaohui Yan and Yanyan Lan and Jiafeng Guo, "BTM: topic modeling over short texts," in IEEE Transactions On Knowledge And Data Engineering, IEEE, 2014, pp. 2928-2941.

[12] Sergei Koltcov and Vera Ignatenko and Olessia Koltsova, "Estimating Topic Modeling Performance with Sharma – Mittal Entropy," in Entropy, Multi Digital Publishing Institute, 2019.

[13] Leila Golshani and Einollah Pasha and Gholamhossein Yari, "Some properties of Renyi entropy and Renyi entropy rate," in Information Sciences, Elsevier, 2009, pp. 2426-2433.

[14] Christian Beck, "Generalised Information and Entropy Measures in Physics," in Contemporary Physics, Taylor & Francis, 2009, pp. 495-510.

[15] Sergei Koltcov, "Application of Rényi and Tsallis entropies to topic modeling optimization," in Physica A: Statistical Mechanics and its Applications, Elsevier, 2018, pp. 1192-1204.

[16] S. N. Koltcov, "A thermodynamic approach to selecting a number of clusters based on topic modeling," in Technical Physics Letters, Springer, 2017, pp. 584-586.

[17] Yu L Klimontovich, "Problems in the statistical theory of open systems: Criteria for the relative degree of order in self-organization processes," in Soviet Physics Uspekhi, IOP Publishing, 1989.

[18] Juan M Banda and Ramya Tekumalla and Guanyu Wang and Jingyuan Yu and Tuo Liu and Yuning Ding and Ekaterina Artemova and Elena Tutubalina and Gerardo Chowell, "A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration," in Epidemiologia, Multidisciplinary Digital Publishing, 2021, pp. 315-324.

[19] Gareth James and Daniela Witten and Trevor Hastie and Robert Tibshirani, "Statistical learning," in An introduction to statistical learning, Springer, 2013, pp. 15-57.

[20] Jia Xue and Junxiang Chen and Chen Chen and Chengda Zheng and Sijia Li and Tingshao Zhu "Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter," in PloS one, Library of Science San Francisco, 2020.