Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4670

Enhancing Contributions to Brazilian Social Media Analysis Based on Topic Modeling with Native BERT Models

Giordanno Brunno Bergamini Gomes [**Universidade Estadual de Campinas** | ***bergaminigomes@gmail.com***]
Romis Attux [**Universidade Estadual de Campinas** | ***attux@unicamp.br***]
Cristiano Cordeiro Cruz [**Instituto Mauá de Tecnologia** | ***cristianoccruz@yahoo.com.br***]

Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Av. Albert Einstein – 400. Cidade Universitária Zeferino Vaz. Barão Geraldo – Campinas – SP – Brasil. CEP: 13083-852.

Received: 5 July 2024 • **Published:** 23 July 2025

Abstract This study introduces a computational approach utilizing natural language processing for text analysis, particularly focusing on topic modeling from large-scale textual data. Given the increasing volume of information shared on social media platforms like X (Twitter), there is a pressing need for effective methods to extract and understand the underlying topics in these texts. Continuing the previous work with LDA, BTM, NMF, and BERTopic, we conducted experiments using advanced BERT embedding models tailored for Brazilian Portuguese, namely BERTimbau and BERTweet.BR, alongside the standard multilingual BERTopic model. We also perfomed experiments with LLM embedding models inside the BERTopic structure, NV-Embed-v2, and gte-Qwen2-7B-instruct. Our findings reveal that the gte-Qwen2-7B-instruct model outperforms the others regarding topic coherence, followed by NV-Embed-v2, BERTimbau Large, BERTimbau Base, BERTweet.BR, and the standard multilingual BERTopic. In the case of the BERT models, this demonstrates the superior capability of models trained specifically on Brazilian Portuguese data in capturing the nuances of the language. In the case of the LLM models, the multilingual capability (including Portuguese) demonstrates the performance of gte-Qwen2-7B-instruct over NV-Embed-v2. The enhanced performance of the gte-Qwen2-7B-instruct model highlights the importance of larger model sizes in achieving higher accuracy and coherence in topic modeling tasks. These results contribute valuable insights for future research in social, political, and economic issue analysis through social media data.

Keywords: Computational Social Sciences, Digital Humanities, Natural Language Processing, Social Media Analysis, Topic Modeling

1 Introduction

Although Digital Humanities — characterized by the intersection between computing and the humanities — is consolidating as a research field [Guimarães *et al.*, 2023] and Computational Social Sciences has experienced a significant increase in prominence over the past decade, there is still much to contribute to these fields due to their recent emergence and the nascent nature of some of their institutional infrastructures [Lazer *et al.*, 2020]. Digital Humanities and Computational Social Sciences are research areas with significant potential for development, especially within the modern confluence of abundant data and the growing availability of high-performance hardware.

One area that brings the humanities and computing closer together is undoubtedly natural language processing (NLP) since texts of all kinds are widely used in the humanities. There are several interesting examples of this confluence [Robila and Robila, 2020], among which we cite, by way of example, the work of Sumikawa *et al.*, 2018. In it, temporal references were extracted from texts on the social network X (Twitter) over 11 months in 2016 and 2017. This extraction was done using a rule-based computational tool called HeidelTime. With these temporal references, it was possible to analyze collective memories [Halbwachs, 1950] on X (Twitter).

This example clearly indicates that topic modeling has emerged organically as a way of determining the subjects discussed in large sets of texts or in large texts. Examples of models are: the simple counting of the total frequency of words in the Bag of Words [Nisha and Kumar R, 2019], LSA (Latent Semantic Analysis) [Deerwester *et al.*, 1990], NMF (Non-Negative Matrix Factorization) [Paatero and Tapper, 1994], pLSA (Probabilistic Latent Semantic Analysis) [Hofmann, 1999], LDA (Latent Dirichlet Allocation) [Blei *et al.*, 2003], Hierarchical Dirichlet Processes [Teh *et al.*, 2004], STM (Structural Topic Model) [Roberts *et al.*, 2013], BTM (Biterm Topic Model) [Yan *et al.*, 2013], W2V-GMM (word2vec Gaussian Mixture Model) [Sridhar, 2015], CorEx (Correlation Explanation) [Gallagher *et al.*, 2017], Top2Vec (Topic to Vectors) [Angelov, 2020] and BERTopic (Topic Bidirectional Encoder Representations from Transformers) [Grootendorst, 2022b].

Social media are sources of great importance for the humanities. One social network that stands out for its wide availability of texts is X (Twitter), which had around 368 million active users in 2022.[1] The platform offers the potential to gather a substantial corpus of data for analysis, encompassing a vast array of subjects and perspectives from a multitude of individuals. Furthermore, the data can be examined at the country and other geographical levels.

[1] https://www.statista.com/statistics/303681/twitter-users-worldwide/

In this work, an extended version of Gomes and Attux, 2023, we present a more comprehensive methodology for analyzing social media, thus seeking to contribute to using NLP in the context of the humanities. In addition, we present a data collection methodology specific to Brazil. We also aim to compare models from different approaches (probabilistic, matrix, short text expert, deep learning neural) for modeling topics in these media and to evaluate the best performance for this task. In employing a probabilistic approach, we utilize LDA; in utilizing a matrix approach, we employ NMF; in utilizing a specialized short text methodology, we employ BTM; and in utilizing a neural approach, we employ BERTopic with diverse embedding models.

Moreover, as an extension of the investigation presented in Gomes and Attux, 2023, we added experiments with BERT models trained in Brazilian Portuguese as embeddings for BERTopic, BERTimbau, and BERTweet.BR, to evaluate whether their performance can be superior when modeling tweet topics in the same language in which they were trained. Additionally, we incorporated embeddings from Large Language Models (LLMs), a recent class of models comprising billions of parameters that achieve state-of-the-art performance in NLP tasks. Specifically, we utilized the NV-Embed-v2 and gte-Qwen2-7B-instruct as embedding models in BERTopic.

The paper is structured as follows: Section 2 covers a review of related work, Section 3 provides a brief review of the models used, Section 4 presents the proposed methodology, the results are in Section 5, and the conclusions are set out in Section 6.

2 Related Work

Traditional opinion polls have long connected public opinion and politicians during elections, but they are often limited, costly, and time-consuming, especially when addressing economic issues. In recent years, social media platforms like X (Twitter) have provided an alternative by enabling the collection of large-scale public opinion data. Karami *et al.*, 2018 introduced a computational approach for public opinion mining that explores discussions of economic issues during elections by combining sentiment analysis and topic modeling with LDA. Applied to millions of tweets, this method effectively analyzed public economic concerns during the 2012 US presidential election, offering a more efficient means of gauging public sentiment.

In another study, Xue *et al.*, 2020 also utilized LDA to examine COVID-19-related discussions on X (Twitter). By identifying dominant topics and themes, the researchers captured the evolving nature of public sentiment and the key areas of interest among X (Twitter) users as the pandemic unfolded.

Boon-Itt and Skunkan, 2020 investigated public awareness and concerns regarding COVID-19 by analyzing 107,990 English-language tweets posted between December 13, 2019 and March 9, 2020. Using keyword frequency, sentiment analysis, and topic modeling with the LDA algorithm, the researchers identified key themes in the discussions. The results revealed three main aspects: the spread and symptoms of COVID-19, which progressed through distinct stages; a generally negative public sentiment toward the pandemic; and three thematic categories—pandemic emergency, control measures, and media reports on COVID-19. The findings highlight X (Twitter) as an effective platform for gauging public awareness and sentiment, offering insights that can inform health departments in addressing public concerns and improving communication during health crises.

Similarly, Lyu and Luli, 2021 employed LDA to identify emerging topics in public discourse on X (Twitter) related to COVID-19, focusing on conversations surrounding the Centers for Disease Control and Prevention (CDC) in the United States. Their study revealed critical trends and concerns about public health that were prevalent during the pandemic.

Egger and Yu, 2022 explored the potential of social media data for social science research, particularly through data-driven approaches such as topic modeling, which offer new insights into social phenomena. However, the short and unstructured nature of social media content presents challenges in data collection and analysis. To address this, the study evaluates the performance of four topic modeling techniques — LDA, NMF, Top2Vec, and BERTopic — using X (Twitter) posts as the reference point. The research highlights the strengths and weaknesses of each method in a social science context, concluding that BERTopic and NMF are particularly effective for analyzing X (Twitter) data.

KH *et al.*, 2022 aimed to model and visualize topics discussed on X (Twitter) by the Makassar community in Indonesia. The study utilized the LDA algorithm to identify and display the prominent themes within these discussions. The findings revealed that effectively applying LDA highlighted the most frequently occurring terms and topics, providing insights into the trends and prevalent critical themes in the Makassar community's social media discourse.

Xu *et al.*, 2022 discussed the politicization of public health issues during the COVID-19 pandemic by developing a computational framework to analyze web-based discourse across different user groups. Focusing on mask-wearing discussions, the researchers clustered X (Twitter) users based on their identities and interests using their bios. They employed BERT Topic modeling to identify and track discourse trends over time. The analysis revealed that political groups and the general public discussed both the science of mask-wearing and the partisan politics of mask policies, with populist rhetoric also emerging. Notably, public health professionals participated less in these discussions. The study emphasizes the importance of user classification in understanding the political context of online public health discourse. It demonstrates the effectiveness of BERT Topic modeling in analyzing short social media texts.

Uthirapathy and Sandanam, 2023 examined public perceptions of climate change by analyzing X (Twitter) data using topic modeling and sentiment classification techniques. LDA method was employed to identify topics discussed, while BERT uncased model was used for sentiment classification, labeling sentiments as pro-news, supportive, neutral, or anti-climate change. The performance of the models was evaluated using precision, recall, and accuracy metrics, with the BERT uncased model achieving the best results, outperforming other methods. The study highlights social media as a valuable tool for understanding public opinions on climate change.

Lotto *et al.*, 2023 investigated the prevalence and topics of ”fluoride-free” tweets to better understand public concerns and the spread of misinformation regarding fluoride. Analyzing 21,169 tweets from 2016 to 2022, using LDA, the researchers identified three key themes: promoting a healthy lifestyle, consuming natural and organic fluoride-free products, and recommending fluoride-free alternatives. The study found that tweets discussing ”fluoride-free” decreased between 2016 and 2019 but increased from 2020 onward, likely driven by misinformation about fluoride's effects. The authors highlight the importance of public health authorities and professionals in addressing the spread of such misinformation to mitigate potential health risks.

Ramamoorthy *et al.*, 2024 explored the potential of social media, notably X (Twitter), in addressing diabetes prevention and management. To understand the scope and content of diabetes-related discussions in India, the authors assess and compare various topic modeling techniques, including LDA, NMF, BERTopic, and Top2Vec. X (Twitter) data from November 2022 to February 2023 was analyzed, with NMF outperforming LDA, and BERTopic surpassing Top2Vec. Eight key topics emerged, focusing on diabetes management, awareness, risk factors, diet, and lifestyle changes. Influential users, primarily healthcare professionals and organizations, were identified as crucial information disseminators.

Rao *et al.*, 2024 analyzed the use of social media data from X (formerly Twitter) to analyze language patterns related to drug use, focusing on differentiating between street names and brand names of prescription drugs. To improve classification, the study applied BERTopic along with Uniform Manifold Approximation and Projection (UMAP) and k-means clustering, generating topics from 170,618 tweets with street names and 245,145 tweets with brand names. Sentiment analysis using the Valence Aware Dictionary and Sentiment Reasoner (VADER) showed that both data sets tended toward positive sentiment. Still, brand-name discussions were more consistent in sentiment classification than street-name discussions. Logistic regression models incorporating tweet text improved accuracy by 40%, highlighting the importance of context for understanding sentiment in drug-related tweets. The findings suggest that engagement metrics alone are insufficient for predicting sentiment without the contextual information from tweet text.

This study aimed to develop a more comprehensive methodology based on the related works presented. It includes real data collection, a sampling strategy, an analysis of sample size, and a search for the most suitable model, ranging from classical to more recent approaches, by fine-tuning their hyperparameters. Additionally, the methodology seeks to address gaps identified in prior research, providing a more robust framework for data analysis and model optimization across various contexts.

3 Models

The Latent Dirichlet Allocation model was proposed by Blei *et al.*, 2003 and has become, over the years, one of the best-known and most used models in topic modeling. It corresponds to a generative probabilistic corpus model, in which documents are represented as random mixtures in latent topics, each characterized by a word distribution. Its hyperparameters α and η refer to the Dirichlet distribution. The first describes an a priori assumption on the distribution of document topics. The second represents an a priori assumption on the topic-word distribution. They need to be adjusted, a process described in Section 3.

The LDA model has limitations: according to Jónsson, 2016, LDA is not a model with outstanding performance for modeling topics in short texts. Since one of the aims of this work is to explore X (Twitter), and since it essentially consists of short texts, it is clear that LDA is not the most appropriate model. It was, therefore, necessary to work with a model aimed at short texts.

A proposal along these lines is the Biterm Topic Model, presented by Yan *et al.*, 2013. This model extracts topics by modeling the generation of word co-occurrence patterns. This model was chosen because of its superior performance in Jónsson, 2016.

Non-Negative Matrix Factorization was introduced by Paatero and Tapper, 1994 under the concept of Positive Matrix Factorization, according to Wang and Zhang, 2013. However, the approach was not used for topic modeling, which came later. NMF is a mathematical procedure in which a matrix of non-negative values is decomposed into two new matrices so that the product of these two new matrices is equal to the original matrix [Churchill and Singh, 2022].

The decomposed matrix is the *document-word matrix*, consisting of a set of documents in which a vector represents each document. The two resulting smaller matrices are the *word-topic matrix*, which can be interpreted as the distribution of topics concerning words, and the *document-topic matrix*, which can be interpreted as the distribution of topics about documents [Churchill and Singh, 2022].

As the last model class for experimenting, comparing, and analyzing short texts on X (Twitter), we used a more recent one that uses state-of-the-art methods for natural language processing tasks. That is BERTopic, an extension of Bidirectional Encoder Representations from Transformers (BERT) proposed by Grootendorst, 2022b. BERTopic is a language model based on a pre-trained transformer, which generates document representations. Its standard multilingual embedding model *paraphrase-multilingual-MiniLM-L12-v2* [Reimers and Gurevych, 2019] groups them, and finally creates topic representations with the class-based TF-IDF procedure.

Extending the work presented in [Gomes and Attux, 2023] and enhancing BERTopic with embedding models trained specifically for Brazilian Portuguese, we used two variants of language models: BERTimbau [Souza *et al.*, 2020] and BERTweet.BR[2]. These models were chosen for their ability

[2] https://huggingface.co/melll-uff/bertweetbr

to capture the nuances and particularities of Brazilian Portuguese, providing a robust basis for natural language processing (NLP) tasks.

BERTimbau replicates BERT's architecture and pretraining procedures with some adaptations for Brazilian Portuguese. Specifically, BERTimbau was trained with the brWaC corpus (Brazilian Web as Corpus) [Wagner *et al.*, 2018], an extensive dataset representing the Brazilian web. This corpus includes billions of words extracted from various textual domains and genres, broadly covering the Portuguese language used in Brazil.

BERTimbau is available in two versions: BERTimbau Base, which has 110 million parameters, and BERTimbau Large, with 330 million parameters. The Base version is designed to be more efficient in computing resources, making it easier to integrate into applications requiring less processing power. On the other hand, the Large version, with a more significant number of parameters, can capture more complex linguistic nuances, providing a deeper understanding of Brazilian Portuguese texts. These features make BERTimbau a powerful tool for various NLP applications, from sentiment analysis to topic extraction and text comprehension.

On the other hand, BERTweet.BR is an adaptation of the BERTweet model initially developed by Nguyen *et al.*, 2020 for analyzing English texts on X (Twitter). BERTweet.BR was trained from scratch, following the pre-training procedure of RoBERTa (Robustly Optimized BERT Pretraining Approach) [Liu *et al.*, 2019]. For this, a corpus of approximately 9 GB was used, containing 100 million tweets in Brazilian Portuguese. The adaptation of BERTweet to Brazilian Portuguese allows it to capture the particularities of the informal and dynamic language used on social networks, especially X (Twitter). This capability is crucial for tasks that involve the analysis of short, informal texts, such as sentiment analysis, topic detection, and other applications that require a detailed understanding of the language of social networks.

By integrating BERTimbau and BERTweet.BR with BERTopic, we can identify and extract topics in a possibly more efficient and accurate way in texts written in Brazilian Portuguese. Using these models can help BERTopic deal with the richness and complexity of the Portuguese language, providing more relevant and contextualized results for the Brazilian reality. This integration is particularly useful for various applications, from analyzing large volumes of textual data to understanding communication trends and patterns in different contexts and platforms.

With the rise of Large Language Models (LLMs), leveraging their capabilities for topic modeling is a logical step. To this end, we drew inspiration from the work of Invernici *et al.*, 2024, which employs LLMs for embeddings based on the Massive Text Embedding Benchmark (MTEB) leaderboard [Muennighoff *et al.*, 2022]. Specifically, we used NV-Embed-v2 [Lee *et al.*, 2024] — a general-purpose embedding LLM with 7.85 billion parameters — for this task. This model currently (November 18th, 2024) holds the top position on the Overall MTEB English leaderboard. MTEB is a benchmark that evaluates models across eight embedding tasks, encompassing 58 datasets and 112 languages.

However, since NV-Embed-v2 supports only English, we also utilized gte-Qwen2-7B-instruct [Li *et al.*, 2023], a multilingual embedding model with 7.61 billion parameters based on the Qwen2-7B LLM. It holds the best position on the Overall MTEB French leaderboard for a Portuguese-supported model (November 18th, 2024). The French leaderboard was selected because it does not include Portuguese, and it is the only Latin language (like Portuguese) compared to other languages from the leaderboard (English, Chinese, Polish, and Russian).

4 Methodology

This section presents the methodology used in this work from the point of view of: 1) assessing the quantity of data, 2) collecting and pre-processing the data, and 3) inserting the data into the models for parameter sensitivity and performance analysis. The subsections establish a sequence and detail each of the stages.

4.1 Delimiting the Language of Tweets

The first step in the methodology consisted of assessing how it would be possible to focus our analysis on content produced in the Brazilian context since this focus was a central motivation for the research. Although X (Twitter) has a search engine that includes the location indicated by the person posting a message, we realized that there were a significant number of messages marked as having been posted in Brazil that were not relevant to us, such as messages posted by visitors attracted by events like the 2014 World Cup or the 2016 Olympics.

We considered that we could use the Portuguese language as a sieve since we had the perception that the number of Brazilian users would be a large majority in this Portuguese-speaking approach. That was confirmed by the ranking of countries with the largest number of X (Twitter) users presented on the Statista[3] platform and by the fact that the number of Brazilian users (19.05 million) is more than 13 times the number of users from Portugal (1.40 million),[4] which is the country with the third largest number of Portuguese-speaking people in the world.[5] Angola, as the second country with the most Portuguese speakers, had only 71.4 thousand X (Twitter) users in 2022.[6]

4.2 X (Twitter) Data Collection

Sample size is a fundamental issue when collecting data: it is necessary to investigate the number of ”tweets” that can represent the subject you want to study. However, the statistical distribution of the data was unknown to us a priori. There was, therefore, no model to determine a sample size that could guarantee a population for a defined confidence

[3] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[4] https://datareportal.com/reports/digital-2022-portugal#: :text=Numbers%20published%20in%20Twitter’s%20advertising,total%20population%20at%20the%20time.

[5] https://www.worlddata.info/languages/portuguese.php

[6] https://datareportal.com/reports/digital-2022-angola

and margin of error. There is, thus, difficulty in defining the number of tweets that can work reliably.

One strategy presented by Krippendorff, 2018 is to carry out sampling experiments to discover a set size. The first collection adopted in this work consisted of obtaining three samples of 50,000 tweets in the year 2021 and comparing them to analyze the adequacy of the size. The three samples collected were composed of the keyword "communism", i.e. they contained "communism" and/or "#communism". This word was chosen because of its centrality in the far-right discourse, which has occupied a large space in the Brazilian political debate over the last ten years [Machado and Colevati, 2021], and also because it is an example of the application of the methodology developed in this work to a social, political, and economic issue.

To collect the data, we used the X (Twitter) API.[7][8] As it returns a maximum of 500 tweets per response, for each set with this maximum returned, the API was provided with a pseudo-random date and time in the year 2021. However, the maximum of 500 was only sometimes returned, as there were unavailable tweets. Another issue is that X (Twitter) has different types of tweets[9]; they are tweets general, tweets status, retweets, replies, mentions, and comments. The collection covered all these types.

4.3 Preprocessing

Once the sample distributions had been inspected, we preprocessed the tweets' texts as indicated in Allahyari *et al.*, 2017. It consisted of the following steps: 1) treating characters and other processes, 2) breaking the text into individual units (known as tokenizing), 3) removing stop words, 4) lemmatization, 5) removing stop words generated by lemmatization.

In treating characters, we removed non-alphanumeric items, i.e., those other than Latin letters and Indo-Arabic numerals. We removed users mentioned in the text and URLs. We replaced the abbreviations with their corresponding full words using a manual list. After this replacement, along with breaking the texts into individual units (tokenizing), all letters were transformed into lowercase, and punctuation was removed.

Stop words are very frequent in a text and have little semantic value, such as prepositions and conjunctions. They were removed using the list in the NLTK library [Bird *et al.*, 2009] and other words added manually to this list.

Lemmatization consists of transforming words into their simplest form, such as putting them in the singular, masculine, or, in the case of verbs, putting them in the infinitive. It involves grouping the various inflected forms of a word so that they can be analyzed as a single item, taking into account the morphological analysis [Allahyari *et al.*, 2017]. This process was carried out using a trained convolutional neural network model from the spaCy library [Honnibal *et al.*, 2020], the small version of which was used due to its lower computational and memory costs. Finally, at the end of the preprocessing, a further removal of stop words was carried out since the lemmatization generated more of them.

4.4 Sample Size Experiments with LDA

Three experiments were conducted using the LDA model to evaluate the sample size. They are in the following subsections: LDA in Samples of 50,000 tweets, LDA with the Split-Half Technique, and LDA with Sample Size Increase. One reason for working with LDA, besides being the most classic model for topic modeling, is that its implementation is the fastest in terms of code execution time. It is called *LdaMulticore* and belongs to the Gensim library [Rehůřek and Sojka, 2010]. Therefore, all the sample size experiments with LDA were carried out with the configuration of five topics, α and η symmetric according to equation 1.

4.4.1 LDA in Samples of 50 Thousand Tweets

Once the bag of words was analyzed, we applied the LDA model with five topics to each of the three samples to compare their results and analyze the representativeness of the sample size. However, comparing topics generated by different LDA models is complex due to the following factors:

- the stochastic nature of the model, requiring each model to be run several times and averages to be compared;
- the unsupervised nature of the model, meaning that there are no expected generated topics, and an appropriate metric is needed to analyze performance;
- the need to choose its hyperparameters to obtain optimal generated topics;
- the need for an expert on the topic of the tweets to better compare the topics generated by the models.

Given these complexities, a suitable metric is needed to analyze the model's performance. The metric used was topic coherence C_V, which is the version with the strongest correlation with human evaluations [Röder *et al.*, 2015]. This metric was used first to evaluate the sample size and then also for the performance of the other models (BTM, NMF and BERTopic) that were used to compare all the models.

We used the average topic coherence C_V of 10 runs for each sample to compare these models as a metric. That resulted in a total of 30 runs. We also used the Jaccard distance between the topics generated by the different models in each run and calculated the average distance to evaluate the sample size. This metric was only used in the sample size analysis.

4.4.2 Split-Half Technique

Split-Half is a technique proposed by Krippendorff, 2018, in which samples are divided in half to analyze whether they show the same behavior concerning the whole set. This method aims to ensure the robustness and consistency of the observed patterns across different subsets of the data. By splitting the samples in half, we can compare the behaviors and characteristics of each subgroup to determine if they align with those observed in the complete dataset. We performed this technique on all three samples.

[7] https://developer.twitter.com/en/docs/twitter-api

[8] The API is no longer free as it once was and it was not possible for us to look for a way of creating a public database.

[9] https://help.twitter.com/en/using-twitter/types-of-tweets

After the division, the frequency of the words was calculated in the bag of words. At this stage, the halves showed very similar rankings in order of the most frequent words. That indicates that the behavior was maintained when the sample was split in half. Once the word frequencies had been analyzed, we ran the LDA model 10 times with five topics in each half. Again, we calculated the coherence C_V and the Jaccard distance between the topics generated by the halves.

4.4.3 Sample Size Increase Experiment

As the behavior of the samples changed when we divided them in half, we carried out a sampling experiment also proposed by Krippendorff, 2018, which is to carry out another collection with a different sample size for comparison.

We collected three samples of 100 thousand tweets, twice the initial sample size. The aim was to evaluate the averages of topic coherence C_V and Jaccard distance between topics and analyze whether this sample size would change the behavior of these metrics concerning the initial sample size. Ten runs were also carried out for each sample.

4.5 Selection of the LDA's Hyperparameters

Once the sample size experiments were finished, the subsequent step was to select the LDA hyperparameters. As the LDA model and its optimization methods have several hyperparameters, it was necessary to adjust them, i.e., to find the values of the hyperparameters that perform best.

Before this tuning, defining which hyperparameters are most relevant to performance is necessary since a search for all of them becomes computationally unfeasible. We then tuned the hyperparameters chosen and worked on the literature [Panichella, 2021]. These are the number of topics, α, and η.

First, we carried out a coarse search, i.e., with large steps of values, to have a wider search space. Then, we performed a fine search, i.e., with small steps of topic numbers, while the values of α and η were fixed. Based on Panichella, 2021, the hyperparameter values in the coarse search were:

- number of topics (k): 2, 42, 82, 122, 162
- α: 'symmetric', 'asymmetric', 0.01, 0.1, 1, 10
- η: 'symmetric', 'auto', 0.1, 1, 10

Where 'symmetric' is equivalent to the inverse of k, 'asymmetric' is equivalent to the inverse of the sum of the topic index and the square root of k and 'auto' is an automatic asymmetric learning of the corpus. These equivalences are presented in equations 1 and 2:

$$symmetric = \frac{1}{k}, \tag{1}$$

$$asymmetric(t) = \frac{1}{t + \sqrt{k}}, \tag{2}$$

where t is the topic index, taking values from 0 to $k - 1$. Once the coarse search was complete, the hyperparameter values in the fine search were:

- number of topics: [2, 4, 6, ..., 158, 160, 162];
- α and η with the highest C_V topic coherence values found in the coarse search, as well as combinations proposed in the literature.

4.6 Application of BTM and NMF to the Samples and Selection of their Hyperparameters

After the LDA experiments, we tuned the BTM model for the three samples. The reason for choosing this model is that it performed best in Jónsson, 2016. In this process, it was also necessary to define which hyperparameters to look for. For this question, we used Jónsson, 2016 as a reference, whose hyperparameters were: number of topics, α, and β (equivalent to η).

The BTM package used was bitermplus,[10] which implements Yan *et al.*, 2013 in Cython. In addition, the combinations of values tested also followed Jónsson, 2016:

- number of topics (k): [10 50 100 200];
- α: [$1/k$ $50/k$ $100/k$];
- β: [0.001 0.01 0.5].

After the BTM experiments, we chose the hyperparameters of the last model, the NMF. The choice consisted of searching for the number of topics with the highest topic coherence C_V for each of the three samples. The NMF implementation used was *nmf*, present in the Gensim library [Řehůřek and Sojka, 2010].

4.7 Application of BERTopic with paraphrase-multilingual-MiniLM-L12-v2, BERTimbau, BERTweet.BR, NV-Embed-v2 and gte-Qwen2-7B-instruct

With BERTopic, we applied topic modeling to the samples. In search of several topics with greater topic coherence C_V, we reduced this number using the implementation of the *reduce_topics* function, which recursively merges pairs of topics from sample data using Agglomerative Clustering [Murtagh and Legendre, 2014]. [11] The implementation present in the Grootendorst, 2022a repository was used to apply the BERTopic model in its default configuration, multilingual embedding model *paraphrase-multilingual-MiniLM-L12-v2*.

In this extended work, BERTopic with the multilingual embedding model *paraphrase-multilingual-MiniLM-L12-v2* was re-run to compare the new models.[12] It was then reapplied to the three samples, and their number of topics was reduced in search of greater topic coherence C_V.

[10] https://bitermplus.readthedocs.io/en/stable/index.html

[11] Agglomerative Clustering has been implemented since version 0.14.0 of the BERTopic package. Before version 0.14.0 of the BERTopic package, the number of topics was reduced by calculating the c-TF-IDF matrix of the documents and then iteratively merging the least frequent topics with the most similar based on their c-TF-IDF matrices.

[12] The version of the BERTopic package in this extended work was 0.16.2, i.e. with topic reduction by Agglomerative Clustering. While the version of Gomes and Attux, 2023 was 0.11.0, i.e. with topic reduction by the c-TF-IDF matrix. Therefore, the results of topic coherence C_V change.

Using the modular structure of the BERTopic package, which allows each step to be replaced by a different procedure, we explored different embedding models to optimize the coherence of the topics generated. BERTopic's flexibility allowed us to experiment with various approaches and adapt the topic modeling process to the specifics of the data. First, we used BERTimbau in its two sizes, Base and Large, as the embedding model, on the same three samples of 50 thousand tweets. The same procedure was then carried out by swapping BERTimbau for BERTweet.BR, NV-Embed-v2, and gte-Qwen2-7B-instruct as embedding models. For each of them, the number of topics was reduced to find the one with the highest topic coherence C_V.

5 Results and Discussions

In **Table 1**, you can see a summary of the experiments with different sample sizes, i.e., variations in the number of tweets collected. We chose to work with 50 thousand tweets in the samples given the compromise between better C_V/d_J ratio performance and a medium computational cost compared to the quantities of 100 thousand and 25 thousand tweets.

Table 2 presents a summary of the performances achieved in the experiments to choose the hyperparameters for each model. According to the experiments from Gomes and Attux, 2023, BERTopic Multilanguage[13] performed better than NMF, BTM, and LDA. That shows that this neural model for topic modeling has an advantage over the older models, LDA and BTM (which are based more strictly on probabilistic tools) and NMF (based on linear algebra techniques).

Furthermore, BERTopic could have been employed without the need for pre-processing, because it is a model capable of considering the context of a text. The model can weigh word order, stop words, and punctuation. In this instance, the model would likely perform even more effectively than the other models. Nevertheless, we maintained uniform methodology across all models to facilitate comparison.

Table 2 also summarizes the performances achieved in the new experiments carried out for each model. These experiments revealed that BERTimbau Large obtained the best performance among BERT models, followed by BERTimbau Base, BERTweet.BR, and, lastly, BERTopic with multilingual standard embedding. These results highlight the superiority of embedding models trained specifically for Brazilian Portuguese in modeling tweets topics in this language. The linguistic specialization of the BERTimbau and BERTweet.BR models, which were trained with Brazilian Portuguese corpora, allows them to more accurately capture the nuances and variations of the language, resulting in better performance in topic generation.

Moreover, the fact that BERTimbau Large outperforms BERTimbau Base demonstrates the importance of the number of parameters in language models. BERTimbau Large, with 330 million parameters, has a greater capacity to capture and represent complex semantic information than BERTimbau Base, which has 110 million parameters. This expansion in model capacity translates into improved performance in topic modeling, thereby demonstrating that larger models, with a greater number of parameters, exhibit enhanced capacity to comprehend and process texts. That is particularly pertinent for applications that necessitate a profound and intricate understanding of the text, indicating that investing in larger models can yield benefits regarding the quality of the outcomes obtained.

Nevertheless, with the inclusion of LLMs, BERTimbau Large has been surpassed. NV-Embed-v2, being 23 times larger, achieved a higher average topic coherence score C_V. Moreover, gte-Qwen2-7B-instruct demonstrated even better performance due to its support for Portuguese. Additionally, these LLMs outperformed the topic coherence C_V achieved by recent works such as Urhan, 2024, Goswami *et al.*, 2024, and Shyu and Weng, 2024. This improvement highlights the significant advancements in topic modeling enabled by the latest LLM-based embedding techniques, particularly in multilingual.

6 Conclusion

This paper presents a methodology for performing social media analysis based on topic modeling. Firstly, we defined a way of specifying the data for Brazil based on language. Secondly, we presented a procedure for collection and pre-processing. Thirdly, we defined a sample size analysis method to determine the number of collections that can be representative of a study. Lastly, elements were indicated to improve the models' performance by their hyperparameters.

Based on the results, we carried out a comparative analysis and observed that the most recent model, which is characterized as a deep learning neural strategy in the category of pre-trained transformers, BERTopic with NV-Embed-v2 and gte-Qwen2-7B-instruct as LLM embedding models, performed better. The advantage of models not based on a bag of words [Shadrova, 2021], as is the case with BERT e LLM models, which create representations of texts in multidimensional vector space, taking the entire context into account and then group the texts into topics.

Another essential point to note is the influence of the models' size on their performance. In our experiments, gte-Qwen2-7B-instruct, which has a significantly higher number of parameters compared to the other models and Portuguese support, showed superior performance in terms of topic coherence. This result is unsurprising, as larger models have a greater capacity for learning and representation, allowing them to capture more detailed and subtle nuances in the text. With 7.61 billion parameters, gte-Qwen2-7B-instruct can create embeddings of words and phrases that more accurately reflect the complex semantic relationships in textual data, resulting in more coherent topics. Despite the 240 million fewer parameters of gte-Qwen2-7B-instruct than NV-Embed-v2, the former has demonstrated superior performance, primarily due to its compatibility with the Portuguese.

Furthermore, gte-Qwen2-7B-instruct's superiority in topic coherence underscores the importance of model size in

[13] The results presented for BERTopic Multilanguage, BERTimbau and BERTweet.BR refer to applications of the models with version 0.16.2 of the package. In [Gomes and Attux, 2023] it was run on version 0.11.0, which had a different topic reduction method as explained above in Section 3.6.

Table 1. Total averages of topic coherence C_V and Jaccard's distance for the three different sample sizes.

Metrics	25,000 tweets	50,000 tweets	100,000 tweets
Topic Coherence C_V	0.275	0.299	0.235
Jaccard's Distance	0.837	0.884	0.842
C_V/d_J	0.329	0.338	0.279

Table 2. Maximum topic coherence values C_V for each sample in each model.

Sample	LDA	BTM	NMF	BERTopic Multilanguage	BERTimbau Base	BERTimbau Large	BERTweet.BR	NV-Embed-v2	gte-Qwen2-7B-instruct
1	0.594	0.624	0.651	0.758	0.767	0.772	0.768	0.769	0.767
2	0.540	0.592	0.641	0.738	0.749	0.749	0.755	0.754	0.761
3	0.561	0.610	0.662	0.750	0.758	0.759	0.748	0.759	0.760
Mean	0.565	0.609	0.651	0.749	0.758	0.760	0.757	0.761	0.763

topic modeling. A larger number of parameters allows the model to develop a richer and more detailed representation, which is crucial for creating relevant and consistent topics. This aspect is essential when dealing with texts in Brazilian Portuguese, where larger models can better capture linguistic and contextual particularities. Thus, gte-Qwen2-7B-instruct's improved ability to create efficient representations not only demonstrates its superior effectiveness but also justifies investing in larger models for complex natural language processing tasks.

Possible prospects and future work lie in analyzing the topics produced by the models together with an expert on the subject, especially those generated by gte-Qwen2-7B-instruct, which obtained the highest performance. This collaboration could provide deeper insights into the quality and relevance of the topics identified and guide fine adjustments to the models to improve their performance further. In addition, integrating expert feedback can lead to the development of practical applications in specific areas, such as social media analysis, trend monitoring, and decision support.

Therewithal, gte-Qwen2-7B-instruct has great potential, given the results and discussion of Shadrova, 2021. It would be interesting to use it to analyze other social issues using the methodology of this work. Also, use its dynamic capacity, i.e., over time. Another idea would be to use a summarization technique on the groups of documents found by BERTopic, instead of class TF-IDF. That could bypass the bag of words approach in this final stage of the model.

In addition to future work, we aim to explore additional LLMs and evaluate other components of the BERTopic framework beyond embedding construction. That includes examining dimensionality reduction and clustering steps by experimenting with alternative hyperparameters for standard models, as conducted by Invernici *et al.*, 2024, and testing alternative models. Such exploration may uncover further optimizations and improvements in topic modeling performance, enhancing the adaptability and robustness of BERTopic in diverse contexts.

Finally, another avenue for future work is to explore prompt-based topic modeling using LLMs, as demonstrated by Pham *et al.*, 2024. This approach would be adapted to the context of our research, focusing on short-text social media platforms such as X. By tailoring prompt-based techniques to these environments, it may be possible to improve topic coherence and relevance in analyzing dynamic and concise online discussions.

Acknowledgements

We would like to express our gratitude to Twitter for providing access to the data used in this study. The availability of this dataset was crucial for developing and validating our methodology. We acknowledge that access to real-world data significantly enhances the practical relevance of our research. That allowed us to perform more robust analyses and test our approach under realistic conditions, contributing to our findings' overall strength and applicability.

Funding

This study was funded in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001 and CNPq (Proc. 308425/2023-5).

References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques.

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boon-Itt, S. and Skunkan, Y. (2020). Public perception of the covid-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health Surveill*, 6(4):e21978. DOI: 10.2196/21978.

Churchill, R. and Singh, L. (2022). The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s). DOI: 10.1145/3507900.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407. DOI: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Egger, R. and Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7. DOI: 10.3389/fsoc.2022.886498.

Gallagher, R. J., Reing, K., Kale, D., and Ver Steeg, G. (2017). Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Gomes, G. B. and Attux, R. (2023). Contributions to social media analysis based on topic modelling. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*, pages 113–120, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/kdmile.2023.231795.

Goswami, A., Kumar, A., and Pramod, D. (2024). Bursty event detection model for Twitter. In Devismes, S., Mandal, P. S., Saradhi, V. V., Prasad, B., Molla, A. R., and Sharma, G., editors, *Distributed Computing and Intelligent Technology*, pages 338–355, Cham. Springer Nature Switzerland.

Grootendorst, M. (2022a). Bertopic. `https://github.com/MaartenGr/BERTopic`.

Grootendorst, M. (2022b). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Guimarães, S. A. d. S., Rocha, E. S. S., and Mugnaini, R. (2023). Estudo cientométrico da atividade acadêmica sobre as temáticas de humanidades digitais e big data nas universidades estaduais paulistas. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 28:1–34. DOI: 10.5007/1518-2924.2023.e90566.

Halbwachs, M. (1950). La mémoire collective [la memoria colectiva]. *Paris, Francia: Presses Universitaires de France*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/312624.312649.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. DOI: 10.5281/zenodo.1212303.

Invernici, F., Curati, F., Jakimov, J., Samavi, A., and Bernasconi, A. (2024). Capturing research literature attitude towards sustainable development goals: an LLM-based topic modeling approach.

Jónsson, E. (2016). An evaluation of topic modelling techniques for Twitter.

Karami, A., Bennett, L. S., and He, X. (2018). Mining public opinion about economic issues. *Int. J. Strat. Decis. Sci.*, 9(1):18–28.

KH, M., Zainuddin, H., and Wabula, Y. (2022). Twitter social media conversion topic trending analysis using Latent Dirichlet Allocation algorithm. *Journal of Applied Engineering and Technological Science (JAETS)*, 4(1):390–399. DOI: 10.37385/jaets.v4i1.1143.

Krippendorff, K. (2018). *Content analysis*. SAGE Publications, Thousand Oaks, CA, 4 edition.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., and Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062. DOI: 10.1126/science.aaz8170.

Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., and Ping, W. (2024). NV-Embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach.

Lotto, M., Zakir Hussain, I., Kaur, J., Butt, Z. A., Cruvinel, T., and Morita, P. P. (2023). Analysis of fluoride-free content on Twitter: Topic modeling study. *J Med Internet Res*, 25:e44586. DOI: 10.2196/44586.

Lyu, J. C. and Luli, G. K. (2021). Understanding the public discussion about the centers for disease control and prevention during the covid-19 pandemic using Twitter data: Text mining analysis study. *J Med Internet Res*, 23(2):e25108. DOI: 10.2196/25108.

Machado, M. G. and Colevati, J. (2021). Anticomunismo e Gramscismo Cultural no Brasil. *Revista Aurora*, 14(Edição Especial):23–34. Number: Edição Especial. DOI: 10.36311/1982-8004.2021.v14esp.p23-34.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2022). MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316*. DOI: 10.48550/ARXIV.2210.07316.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31:274–295.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English tweets. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-demos.2.

Nisha and Kumar R, D. A. (2019). Implementation on text classification using bag of words model. *SSRN Electron. J.*

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126. DOI:

https://doi.org/10.1002/env.3170050203.

Panichella, A. (2021). A systematic comparison of search-based approaches for LDA hyperparameter tuning. *Information and Software Technology*, 130:106411. DOI: https://doi.org/10.1016/j.infsof.2020.106411.

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., and Iyyer, M. (2024). TopicGPT: A prompt-based topic modeling framework.

Ramamoorthy, T., Kulothungan, V., and Mappillairaju, B. (2024). Topic modeling and social network analysis approach to explore diabetes discourse on twitter in india. *Frontiers in Artificial Intelligence*, 7. DOI: 10.3389/frai.2024.1329185.

Rao, V. K., Valdez, D., Muralidharan, R., Agley, J., Eddens, K. S., Dendukuri, A., Panth, V., and Parker, M. A. (2024). Digital epidemiology of prescription drug references on x (formerly twitter): Neural network topic modeling and sentiment analysis. *J Med Internet Res*, 26:e57885. DOI: 10.2196/57885.

Rehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. pages 45–50. DOI: 10.13140/2.1.2393.1847.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., *et al.* (2013). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Robila, M. and Robila, S. A. (2020). Applications of artificial intelligence methodologies to behavioral and social sciences. *Journal of Child and Family Studies*, 29(10):2954–2966.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2684822.2685324.

Shadrova, A. (2021). Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, 2021. DOI: 10.46298/jdmdh.7595.

Shyu, R. and Weng, C. (2024). Enabling semantic topic modeling on Twitter using MetaMap. *AMIA Summits Transl. Sci. Proc.*, 2024:670–678.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Sridhar, V. K. R. (2015). Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 192–200.

Sumikawa, Y., Jatowt, A., and Düring, M. (2018). Digital history meets microblogging: Analyzing collective memories in Twitter. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 213–222, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3197026.3197057.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2004). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in neural information processing systems*, 17.

Urhan, C. (2024). *Enhancing Semantic Understanding by Bridging Topic Modeling and Thematic Analysis: An Empirical Study on Self-Help Twitter Corpus and In-Depth Interviews*, pages 53–71. Springer Nature Switzerland, Cham. DOI: 10.1007/978-3-031-48941-9$_5$.

Uthirapathy, S. E. and Sandanam, D. (2023). Topic modelling and opinion analysis on climate change Twitter data using LDA and BERT model. *Procedia Computer Science*, 218:908–917. International Conference on Machine Learning and Data Engineering. DOI: https://doi.org/10.1016/j.procs.2023.01.071.

Wagner, J., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese.

Wang, Y.-X. and Zhang, Y.-J. (2013). Nonnegative Matrix Factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353. DOI: 10.1109/TKDE.2012.51.

Xu, W. W., Tshimula, J. M., Dubé, È., Graham, J. E., Greyson, D., MacDonald, N. E., and Meyer, S. B. (2022). Unmasking the Twitter discourses on masks during the covid-19 pandemic: User cluster–based BERT topic modeling approach. *JMIR Infodemiology*, 2(2):e41198. DOI: 10.2196/41198.

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., and Zhu, T. (2020). Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. *J Med Internet Res*, 22(11):e20550. DOI: 10.2196/20550.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A Biterm Topic Model for short texts. WWW '13, page 1445–1456, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2488388.2488514.