The effect of political polarization on social distance stances in the Brazilian COVID-19 scenario

Régis Ebeling, Carlos Abel Córdova Sáenz, Jeferson Nobre, Karin Becker

Universidade Federal do Rio Grande do Sul, Brazil
{rebeling, cacsaenz, jcnobre, karin.becker}@inf.ufrgs.br

Abstract. The COVID-19 pandemic changed the routine and concerns of people around the world since 2020. The alarming contagious rate and the lack of treatment or vaccine evoked different reactions to controlling and mitigating the virus's contagious. In this paper, we developed a case study on the Brazilian COVID scenario, investigating the influence of the political polarization in the pro/against stances of social isolation, represented in Twitter by two groups referred to as the Cloroquiners and Quarenteners. We analyzed these groups according to multiple dimensions: a) concerns expressed by each group and main arguments representing each stance; b) techniques to automatically infer from users political orientation, c) network analysis and community detection to characterize their behavior as a social network group and d) analysis of linguistic characteristics to identify psychological aspects. We propose combining two topic modeling techniques, LDA and BERTopics, to understand each stance's concerns in different granularity levels. Our main findings confirm that Cloroquiners are right-wing partisans, whereas Quarenteners are more related to the left-wing. Cloroquiners and Quarenteners' political polarization influences the arguments of economy and life and a stronger support/opposition to the president. As a group, the network of Cloroquiners is more closed and connected, and Quarenteners have a more diverse political engagement with a community of users polarized only with left-wing politicians and his supporters. In terms of psychological aspects, polarized groups come together on cognitive issues and negative emotions.

Categories and Subject Descriptors: H.4 [**Social and Behavioral Sciences**]: General; I.7 [**Document and Text Processing**]: General

Keywords: political polarization, COVID-19, analysis framework, group behavior

1. INTRODUCTION

Brazil has experienced an increasingly politically polarized scenario in the last years. Since the 2014 highly competitive Presidential elections, which resulted in an impeachment two years later, voters have strengthened their political beliefs regarding left or right. The 2018 Presidential elections - won by Jair Bolsonaro - further divided the population when voters chose sides mostly base on "anti" stances, mainly against the Labour party, which has ruled the country since 2002. Bolsonaro's voters were motivated by a liberal economy, a more conservative agenda, and the fight against corruption. In many aspects, the situation is analogous to the United States of America (US) regarding Donald Trump's victory. The onset of SARS-Cov-2 and the related Coronavirus Disease 2019 (COVID-19) arrived in this scenario, and unfortunately, mixed attitudes undermined measures to handle the pandemic. At the time of writing, Brazil has consistently ranked for months in the Top 3 lists of the number of cases and deaths.

The proper actions to mitigate and control the COVID-19 pandemic in Brazil have been discussed in this highly politically polarized scenario. As in many countries, the dilemma between lives and economy has divided opinions. By March 2020, the Brazilian government's initial response, represented by the then Ministry of Health, Luiz Mandetta, was centered on social isolation, supported by the

This research initiative is supported by FAPERGS (project 19/2551-0001862-2).

scientific evidence available and recent experiences from other countries. This direction was followed by most governors, who had to directly deal with the practical aspects of available hospital beds and resources for medical care, considering the growing number of cases. On the other hand, President Bolsonaro has consistently questioned the impact of social isolation on the economy, saying that it was most harmful to the population than the virus itself. Also, he defended medications with questionable efficacy, such as (hydroxy)chloroquine and the benefits of herd immunity through uncontrolled infection. This clash led to the resignation of two Ministers of Health with a medical background and their replacement by an interim minister with military training. Two major pro/anti-social distance movements, commonly referred to the traditional media and social networks as "*Quarenteners*" and "*Cloroquiners*", represent this polarized scenario. In social networks, the views of the two groups were expressed massively, using hashtags such as *#OBrasilNãoPodeParar* and *#OBrasilTemQuePararBolsonaro* (BrazilCannotStop and BrazilMustStopBolsonaro).

Based on data collected on polls, studies have demonstrated that political polarization has influenced the population's behavior towards COVID-19 in the US [Makridis and Rothwell 2020; Bruine de Bruin et al. 2020]. In addition, there are a few peer-reviewed published articles that investigate the perception about COVID-19 on data extracted from social media. Most of them cover topic modeling and diffusion models considering the pandemic [Ordun et al. 2020]. Some initiatives discuss the relation of topics with political polarization: [Jiang et al. 2020] examines geographic differences in online discourses, [Sha et al. 2020] analyzes narratives according to governmental decision making, and [Rao et al. 2020] investigates the relationship between political partisanship and anti-science behavior. Linguistic properties have been explored to characterize the political polarization in social media posts in topics varying from mass shootings [Demszky et al. 2019] to racial violence [De Choudhury et al. 2016].

In this article, we analyze how political polarization affects the behavior of Brazilian groups with opposite stances regarding social distance as a mitigation measure against COVID-19. Two groups of Twitter users, namely the Cloroquiners and the Quarenteners, represent pro/anti-social distance stances, and their behavior is compared to a group allegedly without political motivations. We propose a multi-dimensional analysis framework that includes: a) techniques to automatically infer the political orientation of Twitter users; b) topic modeling to discover the concerns expressed by each group; c) network analysis and community detection to characterize the political influence in the social network group, and d) analysis of linguistic characteristics to identify the underlying psychological aspects. The analysis aims at answering the following research questions:

- Q1: Do these groups have different concerns?
- Q2: Are these groups politically polarized?
- Q3: Does the polarization affect their social network structure?
- Q4: Do the groups have different psychological aspects?

This article is an extension of our previously presented work [Ebeling et al. 2020a]. We have expanded the analysis of the concerns expressed by these two groups by combining the probabilistic topic modeling method LDA [Blei et al. 2003] with BERTopic [Grootendorst 2020], which leverages joint document and word semantic embedding to find topic vectors. We also provided more details on the analysis of the polarized communities in the network analysis. Finally, we also expanded the theoretical background and related work. In a subsequent work [Ebeling et al. 2020b], we adopted the Identity Protective theory to explain the behavior of Cloroquiners and Quarenteners.

The main contributions of this article are:

- the analysis of two groups with opposite stances regarding social distance in the context of COVID-19, and the influence of political polarization in these stances;

- a multi-dimensional analysis framework that encompasses expressed concerns, political polarization, and social network structure, and psychological aspects;
- a method that leverages two complementary topic modeling techniques (LDA and BERTopic) to understand the arguments that support the stances pro/against-social distance. While the former finds coarse-grained concerns, the latter enables to identify the arguments used to voice each stance;

Our main findings confirm that the political polarization considering the Cloroquiners and Quarenteners movement influences the arguments of economy and health, solutions for the pandemic scenario, as well as the support/opposition to President Bolsonaro. The combined use of two topic modeling techniques (LDA and BERTopic) enabled us to identify the general concerns, and associate them with representative arguments in a finer-grained analysis. The finer grain BERTopic analysis revealed that all arguments within LDA topics embed a political bias, which was not evident in our previous work [Ebeling et al. 2020a]. The expanded social network structure analysis revealed strongly connected communities in both the Cloroquiners and Quarenteners groups. While the Cloroquiners have fewer and larger communities, the Quarenteners have more small communities. Based on the politicians followed, the Cloroquiners have a very politically engaged community of connected users, influenced by right-wing YouTubers and politicians. On the other hand, the Quarenteners have two polarized communities, with a more diverse political connections pattern, influenced by the press, journalists, politicians, and political activists. Both movements share common traits in terms of psychological aspects, such as cognitive sophistication and negative emotions, showing that despite their different stances, they focus on expressing their discontent and how the government actions affect their particular point of view.

This article is organized as follows. Section 2 details the theoretical background that supports the proposed analysis framework. Section 3 summarizes related work. Section 4 details the data used to represent each group and the techniques proposed to analyze their concerns, characterize their political polarization, examine the effect of polarization in the social network structure, and identify the underlying psychological aspects. Section 5 discusses the results in light of our research questions. Conclusions and future work are outlined in Section 6.

2. THEORETICAL BACKGROUND

2.1 Topic Modeling

We applied Topic Modeling to identify the concerns of each group studied and answer the research question Q1. Topic Modeling [Hornik and Grün 2011] consists of constructing generative probabilistic models under the premise that documents are mixtures of latent semantic topics, represented as distributions of words. Mining a corpus for latent semantic topics implies deriving word distributions and inferring how documents are distributed among topics.

One of the most popular techniques for fitting a topic model from a *corpus* is Latent Dirichlet Allocation (LDA) [Blei et al. 2003]. This unsupervised technique treats each document in the corpus as a mixture of topics, where each topic has a probability of being related to the document. In turn, each topic is composed of a list of words (terms), with the respective probability of being related to the topic. LDA results in topics in which the terms are more likely to co-occur together in documents. Topics may have overlapping words. LDA input is a corpus, and the discovery of the number of topics is a parameter k. The output is a set of k topics, consisting of terms and their respective probability *beta* of belonging to the topic, and a *gamma* likelihood that relates each topic and a corpus document. Typical pre-processing actions over the original corpus may improve the results [Denny and Spirling 2018], such as normalization, removal of stop words and special characters/terms, stemming, etc.

There are a few challenges related to the deployment of LDA in practice. First, one has to assign a meaning to each resulting topic, a subjective task. Another issue is the parameter k, since a large k may result in redundant topics, while a small k may not be enough to group documents according to a meaningful semantic interpretation. Measures for evaluating LDA results based on the distribution of topics (e.g., W-Uniform, W-Vacuous, and D-BGround [AlSumait et al. 2009]) produce useful results but are difficult to be interpreted by humans. Metrics related to the consistency of the terms in each topic have been used to measure the interpretability of each topic, such as purity [Manning et al. 2010], Normalized Mutual Information (NMI) [Estévez et al. 2009] and Coherence Value (CV or C-Value) [Röder et al. 2015]. CV is a broad and complete coherence metric, which aggregates other coherence metrics according to four dimensions: a) the segmentation used to divide a set of words into subsets, b) confirmation measures that mark the agreement of pairs of terms, c) methods to estimate the word probabilities for confirmation measures, calculated in different ways [Douven and Meijs 2007], and d) aggregation methods to consolidate these results in a single metric. It has been used in different works as a reference to find the proper number of topics in LDA (e.g., [Walter and Becker 2018; Vargas-Calderón et al. 2019; Puerari et al. 2020]).

Distributed representations of words and documents as embeddings have gained popularity due to their ability to capture semantics. An embedding is a relatively low-dimensional space into which one can translate high-dimensional vectors, such as words or documents. Word2Vec [Mikolov et al. 2013] and Doc2Vec [Le and Mikolov 2014] are classical unsupervised techniques for extracting embeddings that represent words and documents, respectively. More recent techniques allow the discovery of contextual embeddings to represent language models, such as BERT [Devlin et al. 2019]. Top2Vec [Angelov 2020] is an alternative approach for topic modeling, which leverages joint document and word semantic embedding to find topic vectors. Top2Vec is a framework encompassing algorithms to automatically seek dense topics in a collection of documents, assuming that semantically similar documents form topics within the input collection.

The first step to use Top2Vec is to convert all documents in the corpus into semantic vectors using some embedding model to make documents semantically similar in the vector space. The next step aims to reduce the dimensionality of the document vectors since vectors in high-dimensional spaces tend to be very sparse. Top2Vec adopts UMAP [Narayan et al. 2020], a fast and scalable dimensions reduction technique that preserves the data global structure. The final step is to group semantically similar documents by searching for dense areas in the vector space using a density-based clustering algorithm, namely HDBSCAN [McInnes and Healy 2017]. HDBSCAN handles both noise and variable density clusters, and thus it assigns a label to each dense cluster of document vectors and a noise label to document vectors that are not in a dense cluster. The dense areas of document vectors are used to calculate the topic vectors, and noise documents are discarded. The advantage of Top2Vec compared to LDA is that the former does not require the definition of the number of topics to be discovered nor pre-processing actions over the input corpus. As a drawback, this technique can lead to an excessive number of topics that make the interpretation of the results very difficult.

BERTopic [Grootendorst 2020] is an extension of Top2Vec, which provides broader support to embedding models, including the state-of-the-art BERT embeddings. It also encompasses an additional step in constructing the topics using a TF-IDF approach to characterize the most representative and distinct words. In this work, we used LDA and BERTopic as complementary topic modeling techniques to understand the concerns expressed by groups of users regarding social distance.

2.2 Social Network Analysis

To identify if political polarization affects the social network structure of each group (research question Q3), we applied network analysis techniques over their social network structure. Network Analysis consists of studying the properties and characteristics of networks. A network (or graph) is composed of links called edges that connect a set of nodes [Hansen et al. 2020]. A prevalent type of graph is that of social networks, used in internet social media. In this case, nodes typically represent social entities (generally people) connected by edges representing static (e.g., friendship, follower, subscriber) or dynamic relationships (e.g., respond, mention, like). In Twitter, for example, users and their following static relationships can be abstracted as a set of nodes and directed edges, where the edge source is the follower, and the sink is the followed user.

Topological metrics describing the network can represent various properties that can provide insights into users' nature and behavior in a social network. Some metrics are [Costa et al. 2007]:

- Number of nodes and number of edges, which describe how large the network is.
- Nodes' in-degree and out-degree, which in the context of social networks such as Twitter, represent how many users are followers of a node and how many users are followed by the node, respectively.
- Average degree, which is the average of all nodes' degrees, represents how connected the nodes are in the network.
- Average shortest path, which gives an idea of how close, in average, the nodes are to each other within the network.
- Network diameter is the length of the longest shortest path between any two nodes in the network and provides an intuition on how difficult it can be to reach a node from any other on the network.
- Network's clustering coefficient, representing the probability of finding sub-groups of highly connected nodes within the network.
- Closeness centrality is calculated individually for each node, and it is based on the shortest path between a given node and all the other nodes in the network. Nodes with a high closeness score have the shortest distances to all other nodes. Closeness centrality is a way of detecting nodes that can spread information very efficiently through the network.
- Betweenness centrality of a node is based on the number of shortest paths between all pairs of nodes that pass through it. It is a measure of how important a node can be in enabling the communication between other pairs of nodes, given it is part of the shortest path between them. In other words, nodes with high betweenness may have considerable influence within a network by virtue of their control over information passing between different nodes.

Community detection is a technique frequently used to analyze social networks [Bazzan 2020; Bedi and Sharma 2016; Conover et al. 2011]. Community detection aims at finding groups of nodes (communities) that are highly connected to each other, but loosely connected with nodes from other communities [Fortunato 2010]. In social network analysis, this technique makes it possible to identify users who share the same social patterns within the network and even in social spheres such as politics, as studied in this work. There are several algorithms for the community detection task. One of them is the Louvain Method [Blondel et al. 2008], which is focused on the optimization of the network's modularity. In this work, we used this methods as available in the tool Gephi[1].

2.3 Lexical Dictionaries and Word Functions

To study if the groups have different psychological aspects (research question Q4), we analyzed the functions of the words employed to express concerns, as they reveal emotional and biological states, thinking styles, and other personality traits [Tausczik and Pennebaker 2010]. Lexicons are data structures in which each entry associates a term or phrase to lexical information related to it. A lexicon entry is usually the root of a word, the singular mode of a noun, or the present tense of a verb. For instance, lexicons such as SentiWordNet [Baccianella et al. 2010], NRC [Mohammad and Turney 2013] and LIWC [Pennebaker et al. 2001] have been widely deployed for unsupervised sentiment analysis.

[1]https://gephi.org/

LIWC is a textual analysis tool, which encompasses analysis functionality and a complete lexical dictionary [Pennebaker et al. 2001]. It was conceived from studies and observations that indicated that an individual's writing translated psychological aspects, mainly in illness or recovery processes. Tausczik and Pennebaker highlight two extensive categories of words with different psychometric and psychological properties: content and style [Tausczik and Pennebaker 2010]. From a psychological perspective, content words (e.g., nouns, regular verbs, and many adjectives and adverbs) convey what people say. In contrast, style words (e.g., pronouns, prepositions, articles, conjunctions, auxiliary verbs) reflect how people communicate. The LIWC dictionary construction aimed at automating the word count in the respective psychological categories for the input texts of the tool, and categories can be combined to form a psychological aspect of a person or group. The dictionary categorizes words into categories according to 4 linguistic dimensions: linguistic processes (e.g., pronouns), psychological processes (e.g., social, cognitive), personal concerns (e.g., work, leisure), and informal language (e.g., words of agreement). Categories can be, in turn, divided into subcategories. LIWC enables to characterize of psychological aspects based on the different use of these words. Empirical results using LIWC summarized in [Tausczik and Pennebaker 2010] demonstrate its ability to detect meaning in a wide variety of experimental settings, including attentional focus, emotions, social relationships, thinking styles, and individual differences.

3. RELATED WORK

Twitter has been used to study different social phenomena, such as gender equality [ElSherief et al. 2017], racial equity [De Choudhury et al. 2016], emotional distress due to violent events [Harb et al. 2019; 2020] and political polarization [Conover et al. 2011; Hong and Kim 2016; Garimella and Weber 2017; Preoţiuc-Pietro et al. 2017]. This social network is also used to study and improve processes such as urban mobility [Jerônimo et al. 2017] and customer service by companies [Amora et al. 2018]. Information extracted automatically from the users' profiles allows deepening the understanding of these phenomena, such as the extraction of gender/age from profiles pictures using Face++ [ElSherief et al. 2017], or the prediction of political partisanship based on the polarized users that a user follows [Garimella and Weber 2017]. Political polarization can also be inferred by analyzing clusters of retweets/mentions of tweets from users [Conover et al. 2011] or even by models analyzing a set of features extracted from the tweets [Preoţiuc-Pietro et al. 2017].

LIWC has been deployed to characterize psychological aspects related to political partisanship, such as affect states and pronouns in stances regarding mass shootings [Demszky et al. 2019], honesty and cognitive complexity in USA presidency candidates [Slatcher et al. 2007], negativity, low cognition and low social awareness related to racial violence [De Choudhury et al. 2016], and sentiment and power words for measuring political moderacy in [Preoţiuc-Pietro et al. 2017]. A study on the misperceptions about the COVID pandemics [Pennycook et al. 2020] observed an inverse relationship between cognitive complexity and ideology.

Political polarization and the behavior regarding COVID have been a concern of several works. In the United States of America (US), studies [Makridis and Rothwell 2020; Bruine de Bruin et al. 2020; Milosh et al. 2020; Barrios and Hochberg 2020] reveal that partisan affiliation is often the strongest single predictor of behavior and attitudes about COVID-19, even more powerful than local infection rates or demographic characteristics. A study [Grossman et al. 2020] on mobility data revealed that political partisanship influences citizens' decisions to voluntarily engage in physical distancing in response to communications by their county governor. The influence of political polarization has also been observed in Brazil [Ajzenman et al. 2020; Soares et al. 2021].

COVID-19 in social media is also a very active research area, with many pre-print works. Most of them investigate online conversations in terms of topics, information diffusion, and topics change over time, as summarized in [Ordun et al. 2020]. Regarding political polarization, [Jiang et al. 2020] examines geographic differences in online COVID-19 discourse, relating the polarization to each

US state's political dominance. A longitudinal study [Sha et al. 2020] relates Twitter narratives to Governors and Presidential actions. Another study [Rao et al. 2020] examines the ideological alignment of users along moderacy, political, and science dimensions, concluding that moderacy is the critical influence on the behavior towards science.

Our study differs from related work by examining the political influence of social distancing stances in the Brazilian COVID scenario. We investigated Twitter users according to an analysis framework that combines topic modeling for summarizing concerns, political polarization measuring, social network analysis, and psychological aspects.

4. ANALYSIS FRAMEWORK

To characterize the influence of political polarization in social distance stances in the Brazilian COVID-19 context - as represented by the Cloroquiners and Quarenteners - we propose a multi-dimensional analysis framework to address each research question:

- *Concerns (Q1)*: We used topic modeling to identify the concerns of each group. We propose the complementary use of two distinct techniques, namely LDA and BERTopic. This combination provides a proper balance between the number of topics and interpretability of the results in order to identify the concerns according to distinct granularity levels;
- *Political Polarization (Q2)*: To identify if users of these groups are politically polarized, we proposed an index that measures the level of polarization towards right/left according to the politicians followed;
- *Polarization and Social Network Structure (Q3)*: To analyze if political polarization influences their social network structure, we deployed social network analysis techniques. We analyzed the topological metrics of the respective social networks, identified sub-communities, and analyzed the influential users in polarized communities;
- *Psychological Aspects (Q4)*: we characterized the linguistic styles of each group and analyzed if there are differences in the underlying psychological aspects.

We analyzed the early scenario of COVID-19 (late March, early April of 2020), as represented by campaigns opposing the priority between economy and lives. The remaining of this section describes the data used and details each dimension of the analysis framework.

4.1 Data and Pre-processing

To investigate the political polarization related to pro/against social distance stance in Brazil, we analyzed three groups' characteristics explicitly associated with the social distance in the early scenario of COVID-19 in Brazil. Hashtags represented these groups in the Twitter trending topics by the end of March 2020.

- *Cloroquiners*: to represent the against stance, we selected the hashtag #OBrasilNãoPodeParar (BrazilCannotStop), referring to a federal government's campaign[2] extensively advertised by Bolsonaro himself. The central argument of this campaign is that the economic consequences of social distance are more harmful to Brazilians than the risk of COVID-19 contagious;
- *Quarenteners*: to capture the politically polarized pro stance, we identified the hashtag #OBrasilTemQuePararBolsonaro (BrazilMustStopBolsonaro), created in opposition to the presidential official campaign[3];

[2]https://bit.ly/2AVJJIC
[3]https://getdaytrends.com/trend/%23OBrasilTemQuePararBolsonaro/

Table I. Hashtags and collection numbers per group

Group	Hashtags	Nº Tweets	Nº Users
Cloroquiners	#OBrasilNãoPodeParar	74.395	20.572
Quarenteners	#OBrasilTemQuePararBolsonaro	31.060	10.769
Neutrals	#FicaEmCasa, #FiqueEmCasa	201.499	102.309

- *Neutrals*: we identified the hashtags #FiqueEmCasa and #FicaEmCasa (variations of StayAtHome) to share information related to social distance. Although this group endorses social distance in many respects, we assume it does not encompass any political standpoint. We expect our analysis to confirm this premise.

Table I displays the volume of collected tweets and the respective number of users, considering the identified hashtags (i.e., #OBrasilNãoPodeParar, #OBrasilTemQuePararBolsonaro, #FicaEmCasa, #FiqueEmCasa). We crawled the data using the GetOldTweets API[4], which enables to collect past tweets. We collected tweets posted between March 22 and April 7, when these hashtags were widely used in clear action/reaction in response to the government campaign.

To investigate the influence of bot profiles to boost these hashtags artificially, we deployed the API Botometer[5]. Given a Twitter profile, this API analyzes the account characteristics and returns a probability score related to robot behavior. Due to the high demand received by the API by the time this research was developed, we analyzed samples of randomly selected users: 3,750 Cloroquiners users (18.2%) and 2,792 Quarenteners users (25.9%). We found a similar amount of bots in both samples: 6.46% of Cloroquiner sample users and 5.94% of Quarentener sample users. Considering the API limitations, we also analyzed the number of profiles created within 30 days or less before the hashtag launch. This analysis was motivated by the fact that an account's recency is a strong predictor of robot profiles. We found 5.98% of recent profiles among the Cloroquiners group and 5.54% within the Quarenteners. Thus, based on these two criteria, we conclude there is no significant difference in the use (or potential use) of robots among the users representing these stances. We removed all identified robots and suspended accounts from our analysis.

We applied classical pre-processing actions for all textual analysis, such as normalization and removal of punctuation, special characters, hashtags, mentions, and URLs. We also disregarded all tweets with less than three words.

4.2 Analysis of the Concerns

Identifying the major concerns expressed in each group's tweets is key to understanding the rationale behind the pro/against social distance stances. We propose the combination of two topic modeling approaches: LDA and BERTopic. We regard these techniques as complementary since LDA provides a coarse-grained clustering based on probabilities of words co-occurrence in a documents corpus, and BERTopic helps to identify frequent similar arguments based on tweet similarity. The topics yielded by LDA provide the profile of each group in terms of general concerns. As a complement, BERTopic provides representative arguments used to support their stances and to identify if there is political bias in these arguments. LDA also allows providing context to the large number of topics that result from BERTopic, making their interpretation easier.

The LDA modeling technique allows us to segment each group's tweets into topics that summarize the overall concerns. LDA is associated with two main challenges: parameterization and interpretation of the meaning of the resulting topics. As a coherent division of the documents is dependent on the input parameter k, in our previous work [Ebeling et al. 2020a], we proposed to analyze different outcomes of LDA runs using the CV metric to select an appropriate value for k. We applied the LDA

[4]https://github.com/Jefferson-Henrique/GetOldTweets-python
[5]https://rapidapi.com/OSoMe/api/botometer-pro

to the set of pre-processed tweets corresponding to each group. To find the best k for each of them, we varied k from 1 to 30 and selected the best CV coherence values. Finally, we inspected the results manually for the runs with the best CV values, using the most representative terms for each topic and a random sample of associated tweets. The k chosen in each case represents the smallest set of topics found for a coherent and interpretable set of terms and the least redundancy among them. For Cloroquiners and Quarenteners, $k = 3$ was chosen, and for Neutrals, $k = 8$. The library used was Gensim[6], in addition to using $alpha = 0.5$ and $beta = auto$.

The BERTopic modeling technique transforms the input texts into vectors of BERT embeddings and allows to explore their similarity in the vector space. Thus, this algorithm does not require the specification of the number of topics, as DBSCAN identifies clusters of similar documents according to the dense areas found in this vector space. As a drawback, this approach can lead to an excessive number of topics that make the interpretation of the results very difficult (365 and 155 dense areas of tweets for the Cloroquiners and Quarenteners, respectively). Thus, we propose to combine each technique's strengths by using LDA as a filter to find the least number of coherent topics, and BERTopic to identify the most representative arguments within each LDA topic, and if there is a political bias in the respective arguments.

To facilitate the interpretation of the large number of BERTopics clusters, we propose to compose aggregates for clusters close to each other, referred to as *agglomerations*. To identify agglomerations, we plotted the clusters in a bi-dimensional space using BERTopic library[7], and defined aggregations of overlaid clusters. Then, we analyzed each topic's three largest agglomerations in terms of the number of tweets. Then, we inspected the representative terms and tweets of each agglomeration based on their textual similarity to identify the representative argument. The analysis using BERTopic was developed only on politically polarized groups (i.e., Cloroquiners and Quarenteners) to gain insights into their stances. Table III shows the clusters and aggregates found for each topic of these groups.

As BERT embeddings, we adopted the 'distiluse-base-multilingual-cased'[8], a model that supports the use of 50 different languages. By the time this research was developed, the Portuguese language representation model BERTimbau [Souza et al. 2020] was still unstable, with daily commits, resulting in integration issues with the BERTopic library. In addition, the comparison of these models detailed in [Souza et al. 2020] reveals that this choice delivers good results.

4.3 Political Polarization Index

We propose an index to measure the political polarization of the users according to the right/left politicians they follow. We adopted politicians classified as right/left-wing according to the Ideological GPS[9], a tool that calculates the political orientation of influencers according to patterns found among their followers. We selected the 102 most left-oriented politicians and the 102 most right-oriented ones.

For each user, we collect the list of users followed (*followings*). Then, we calculate the ratio between the number of followed right-oriented politicians and the total number of politicians followed (right or left-oriented). We adopted an offset of 1 for each side to adjust the calculation when a user does not follow right or left politicians. Thus, the value 50% indicates politically neutral users, i.e., either do not follow politicians or follow them in equal amounts. The higher (or lower) the metric value, the more oriented to the right (left) the user is.

Given the computational infrastructure available during the pandemic, we performed this analysis in a random sample of data. We were able to collect the followings list of 8,724 Cloroquiners users

[6]https://pypi.org/project/gensim/
[7]https://pypi.org/project/bertopic/0.3.4/
[8]https://huggingface.co/distilbert-base-multilingual-cased
[9]http://temas.folha.uol.com.br/gps-ideologico/

(42.40%), 4,553 Quarenteners users (42.27%), and 8,361 Neutrals users (8.17%).

4.4 Analysis of the Polarization in the Social Network Structure

The analysis of the social network structure is also based on the users followed. For each group, we constructed a directed graph where the nodes correspond to users from the group or followed by them, and directed edges connect these nodes according to the followings list. Then, we generated topological metrics that characterize each network's complexity (e.g., average degree, average shortest path, diameter, clustering coefficient), as described in Section 2.2.

We used Gephi to identify communities within the social network representing each group. For each community found, we calculated the same graph complexity measures. We also extracted the nodes' centrality metrics, namely proximity centrality (node responsible for directly influencing the network) and betweenness centrality (node responsible for spreading information over the network). We used the same set of politicians to assess the polarization of the resulting communities.

4.5 Psychological Aspects Derived from Linguistic Characteristics

According to Tausczik and Pennebaker [Tausczik and Pennebaker 2010], word functions, represented as LIWC categories, reflect basic psychological states, such as emotional and biological states, thinking styles, honesty, individual differences, etc. We characterized the groups in terms of the following psychological aspects related to LIWC linguistic categories:

- **Cohesion and union:** We investigated if groups have distinctive union/cohesion traits, using word categories that denote group unity. According to [Tausczik and Pennebaker 2010], words from the category *we* can be used to promote group interdependence, and the higher adoption of agreement words (*assent* category) may reveal a greater consensus. Related work has used these word categories to characterize belonging, and involvement [Demszky et al. 2019; De Choudhury et al. 2016].
- **Affect States:** The way people react to traumatic or important events may say a lot about how they cope with the event and the extent to which the event plays a role in the future [Tausczik and Pennebaker 2010]. At the heart of reacting and coping with events is people's emotional response [Harb et al. 2020]. Emotional states' expression constitutes a semantic level relevant to polarization and can assist in detecting ideological levels [Demszky et al. 2019; Preoţiuc-Pietro et al. 2017]. To characterize the emotional state of each group, we used the positive/negative LIWC affect categories (*posemo/negemo*), and the negative sub-categories *anger*, *anxiety* and *sadness*.
- **Cognitive complexity:** According to [Tausczik and Pennebaker 2010], cognitive complexity can be thought of as a richness of two reasoning components: the extent to which someone differentiates between multiple competing solutions and the extent to which someone integrates among solutions. These two processes are captured by the word categories *exclusivity* and *conjunctions*. It is also related to how sophisticated someone's abstract or conceptual thinking is, typically associated with a greater ability to discern between true and false content. *Prepositions* (e.g., to, with, above) and *cognitive mechanisms* (e.g., cause, know, ought) are all indicative of the ability to handle a more complex language. As in [Slatcher et al. 2007; De Choudhury et al. 2016], we adopted the combined use of classes *exclusivity*, *conjunctions*, *prepositions* and *cognitive mechanisms* to characterize the aspects related to cognition revealed by linguistic style.
- **Personal concerns:** To align the topics automatically found from the corpus with the concerns of the individuals in each group, we used LIWC *Personal concerns* subcategories: *work, achievements, money, leisure, home, religiosity, and death.*

We used a Portuguese version of LIWC[10], counting for each tweet the words belonging to all LIWC

[10]http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

Table II. Topics by group

Topics	Tweets	Users	Density	Cloroquiners - Representative Words
0	12855	**5352**	2,40	work, let's, people, want, money, to work
1	11351	4009	2,83	risk, country, quarantine, world, home, hungry
2	**15293**	3984	**3,83**	president, bolsonaro, brazil, everything, speech, God
Topics	**Tweets**	**Users**	**Density**	**Quarenteners - Representative Words**
0	4634	**2159**	2,14	bolsonaro, deaths, government, health, campaign, everybody
1	**5105**	1715	**2,97**	genocide, president, virus, urgent, motorcade, people
2	4799	2027	2,36	brazil, stop, now, let's, bozo, against
Topics	**Tweets**	**Users**	**Density**	**Neutrals - Representative Words**
0	8064	2804	2,80	best, hands, always, water, care, wash
1	10159	2944	3,45	home, stay, can, staying, health, help
2	9851	3114	3,16	distance, social, still, people, virus, corona
3	9755	**5105**	1,91	brazil, covid, bolsonaro, country, cases, deaths
4	12292	3655	3,36	everybody, let's, life, God, pass, love
5	12229	3148	3,88	quarentine, here, do, everything, now, friends
6	**13207**	3159	**4,18**	week, want, night, things, get out, music
7	7930	4309	1,84	today, live, jorge, folks, instagram, congratulations

categories (0 if absent, 1 if present). Then, we analyzed whether there were significant differences in the proportional use of words of LIWC categories in general and the ones related to the four psychological aspects described above. We used the Chi-square test ($alpha = 0.05$) to assess the statistical significance of these differences.

5. RESULTS

5.1 Q1: Do these groups have different concerns?

5.1.1 *LDA Topic Analysis.* The first part of our analysis was to identify the major concerns of each group using LDA. Table II shows the topics found, the number of tweets and users addressing each topic, along with the six (6) most representative words according to the weight to associate with the topic. Then, we conjectured about each topic's central concern based on the most relevant words and the manual inspection of a sample of related tweets.

The Cloroquiners are concerned about the economy, economic consequences of social distance, and politics. Topic 0 addresses the need to get back to work, using terms to denote actions (*let's,now*), subjects (*citizens, people*), and motivation (*work, money, job).* Topic 1 compares social distance in Brazil and the world (*country, world*) and highlights the economic consequences of social distance (*hunger, risk*). Topic 2 expresses support to the president, with mentions (*president*), references to his campaign slogan[11] (*brazil, god*), and his actions (*speech, truth*). Topic 2 encompasses the largest number of tweets, with an average of 3.83 tweets per user, showing Bolsonaro supporters' engagement. Topic 0 includes the largest number of users, with an average of 2.4 tweets per user.

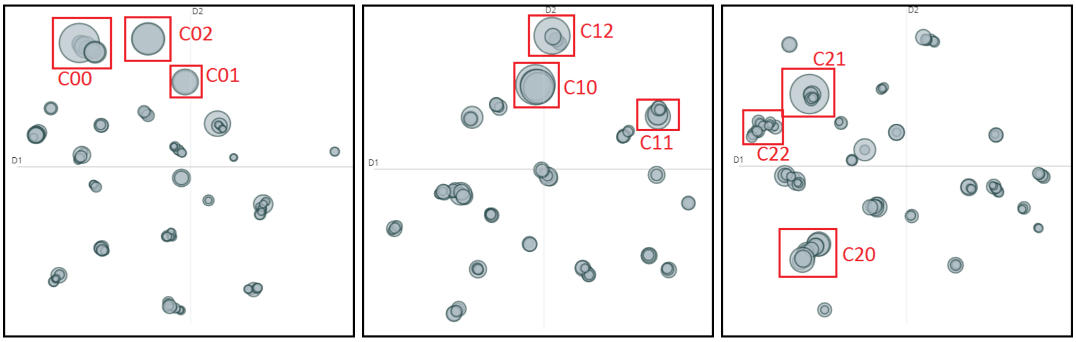
The Quarenteners clearly express their opposition to the government campaign. Topic 0 criticizes the campaign, with mentions of targets (*campaign, government, bolsonaro*), the reason for the criticism (*death*) and mentions of health care professionals (*health*). Topic 1 expresses concerns about the actions of the president and his supporters, with direct mentions (*president*), derogatory adjectives (*genocide*), requests for changes in the government (*urgent*) and criticisms to Bolsonaro's supporters (*motorcade*). Topic 2 emphasizes social distance as the means to stop the spread of the virus, through actions (*stop, let's*), derogatory nicknames (*bozo*) and motivation for social distance (*life*). Topic 1 concentrates the highest number of users, with an average of 2.97 posts per user. The other topics are associated with similar amounts and density of posts (2.14 and 2.36 for topics 0 and 2, respectively).

The Neutrals group is the most diverse one, discussing various aspects related to the virus and social distance. Topic 0 deals with hand washing. Topics 1 and 2 discuss the importance of social

[11] https://bit.ly/2AZpFFc

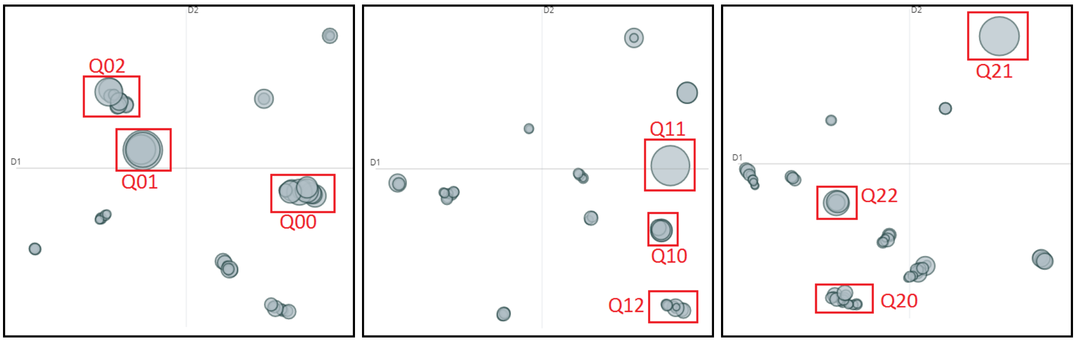
Table III. Topic argument clusters

Group	Topic	#Clusters	#Agglomerations
Cloroquiners	0	127	21
	1	106	18
	2	132	20
Quarenteners	0	55	9
	1	42	11
	2	58	10



(a) Topic 0 (b) Topic 1 (c) Topic 2

Fig. 1. Cloroquiners BERTopic clusters



(a) Topic 0 (b) Topic 1 (c) Topic 2

Fig. 2. Quarenteners BERTopic clusters

distance and the risks of not adopting it. Topic 3 deals with information about pandemics, mainly within Brazil and government actions. Topic 4 concentrates on messages of hope, positivism, and faith. Topics 5 and 6 address social distance consequences, such as daily routines, monotony, and longing for outdoor activities. Topic 7 deals with virtual entertainment. The topics that explicitly address social distance (1, 5, and 6) are the ones that concentrate the highest number of posts, with a density per user of 3.45, 3.88, and 4.18, respectively.

This initial analysis provides evidence of the political engagement of the Quarenteners and Cloroquiners, given that the topics that support/reject the president and his actions (topic 2 for the Cloroquiners and topic 1 for the Quarenteners) concentrate the highest density of posts per user. Users of the Neutrals group do not seem to embed a political bias in expressing their main concerns: the practical impact of social isolation and virtual entertainment.

Table IV. Cloroquiners: Most dense aglomerations

Topic	Agg.	#Clusters	#Tweets	Representative Words	Argument Example
0	C00	11	1247	mild cold, risks, hungry, bums	"If Brazil stops, those who do not die of the virus will die of hunger. Life follows, turn off the television and search the statistics, comparing COVID-19 with other diseases and other causes of death in Brazil and the world, wake up!"
	C01	2	332	work, bills, money, families	"More than wanting to, we need to work! The bills have arrived and the money is running out. Vertical quarantine now"
	C02	1	262	Bolsonaro, president, congratulations, government	"The Brazilian government, under President Jair Bolsonaro, is concerned with millions of Brazilians who depend on their work to support their families and who now because of a political fanfare, causing psychological terror decides to paralyze"
1	C10	4	702	governors, press, hysteria, hygiene	"The president is right, only to keep the elderly and those with chronic diseases in quarantine. Maintain good hygiene rules, avoid crowds and protect your most vulnerable family members"
	C11	9	347	collapse, climbers, politicians, leftists	"They want to break us, they want to see hunger rise but they won't make it"
	C12	7	326	governors, mayors, dictators, fascists	"And they said that the Bolsonaro government would be a dictatorship, they just forgot to say that the dictators would be the governors!"
2	C20	17	852	president, press, Globo, corrupt	"The President is not cold today because he is right! And don't dine but why did the press eat at the fakenews factory"
	C21	9	512	Bolsonaro, reason, traitors, dorianers	"Our President is wonderful! You are truly with the people! He is not hidden in the palace, surrounded by the police like the coward of Doriana and others! How can you not love Bolsonaro?"
	C22	19	378	ministers, military, congratulations, pronouncement	"The best team of ministers this country has ever had! If it weren't for President Bolsonaro, we would never have such a team! Force!"

5.1.2 *BERTopic Analysis.* In this second part of the analysis, we searched for the most representative arguments used by Cloroquiners/Quarenteners within each topic. Table III details the number of clusters per topic and how they were grouped into agglomerations. Figures 1 and 2 display in a bi-dimensional space the distribution of clusters within each topic found for the Cloroquiners/Quarenteners, respectively. They also highlight the three biggest agglomerations identified for each topic. As a convention, the agglomerations are labeled using an acronym that identifies the group, topic, and agglomeration. For instance, the agglomeration C00 denotes the agglomeration 0 for the topic 0 found for the Cloroquiners. Finally, Tables IV and V details each agglomerations analyzed for the Cloroquiners/Quarenteners, respectively. Each table relates the agglomeration to the respective topic, and summarizes the number of aggregated clusters with the respective number of tweets, the top 4 most representative words according to BERTopic, with an example of a representative argument. In the remaining of this subsection, we detail the representative arguments found.

Cloroquiners

Table 1 summarizes the properties of the representative arguments used by the Cloroquinners. Topic 0, which presents arguments and reasons for the population to return to work, has the agglomeration with the largest number of tweets in the Cloroquiners group, C00, representing 9.7% of the tweets in this topic. Agglomeration C00 encompasses 8.6% of the clusters found for this topic, and the central argument is a greater concern with the impact of social isolation on the economic situation (*risks, hungry*). It also minimizes the danger of COVID to people's health (*mild cold*). The second-largest agglomeration, C01, comprises two clusters that argue that the return to work is necessary because the population needs to pay their bills (*bills, money*). The tweets belonging to the third-largest agglomeration, C02, praise the president for the measures taken to prevent a severe economic crisis in the pandemic (*congratulations*). Notice this agglomeration stands for a single cluster and thus to a very dense area of very similar tweets. This more detailed view of the central arguments confirms that Topic 0 expresses the urge to maintain the economy active, considering that the consequences would be more harmful to the population than the virus itself. The support to the president is frequently embedded in the tweets that voice these concerns.

Topic 1 highlights the risks of social isolation to the economics and compares Brazil to the rest of the world. The largest agglomeration (C10) groups tweets expressing the discontent with mayors and governors who adopted social-distance measures to combat the pandemic (*governors*), and with the press that generates excessive panic (*press, hysteria*), considering that good hygiene and vertical isolation would be enough to control COVID contagion (*hygiene*). The second-largest agglomeration, C11, criticizes politicians and left-polarized people (*politicians, leftists*) who want the country to collapse due to unemployment and misery (*collapse*), thus making the president's government even more difficult (*climbers*). They also share opinions of infectologists who endorse vertical isolation, a measure defended by the president. The C12 agglomeration also reports dissatisfaction with the governors and mayors (*governors, mayors*) who are implementing stricter isolation measures, comparing them to dictators who impose rules that hurt the population's freedom (*dictators, fascists*). This finer-grained analysis enables us to understand that the comparison to other countries refers to vertical isolation, a measure with less economic impact adopted by some countries like England, in contrast to horizontal (social) isolation. The arguments of Topic 1 are also intertwined with expressions of support to Bolsonaro's Government and criticisms to all actors (e.g., press, governor, mayors) who undermine his attempts to implement a more flexible model of COVID combat.

The analysis of the aggregations related to Topic 2 confirms the support to the president, his government, and the actions he defends. Recall that this topic has the largest number of tweets and the highest user engagement. The central argument of the largest agglomeration (C20) is criticisms of the opponents of the president, including the press (*press, Globo*), voters, left-wing politicians, or even right-wing opponents. The second-largest agglomeration, C21, expresses support to the president and criticisms to João Dória[12] (*traitors, dorianers*), who adopted isolation-based measures to combat the pandemic in São Paulo state. The third-largest agglomeration, C22, has the support to the president (*congratulations*) as its central argument, praising the president himself, his choice of ministers (*ministers*), the president's pronouncements (*pronouncement*), and even COVID case recovery rates that have been linked to government actions. The analysis of central arguments in this topic revealed that a significant portion of tweets (8.9%) express support to the president through criticisms to a wide range of actors regarded as opponents, from the press to former political allies.

Quarenteners

Table V summarizes the properties of the representative arguments used by the Quarenteners. Topic 0 (criticisms of the government campaign) presents the largest agglomerations in the Quarenteners group. Q00 is the largest agglomeration in this topic, with 657 tweets distributed in 12 clusters (21.8% of the clusters in this topic). These tweets voice the fear of Brazil facing a situation similar to Spain[13] or Italy[14], fear of dying (*die, deaths*) unless social isolation is strictly adopted, and criticisms to the entrepreneurs who lead movements to keep their business open (*economy*). The second-largest agglomeration (Q01) criticizes the president's supporters (*bolsonarists*), stress the growing occupancy

[12]João Doria, governor of São Paulo and a possible candidate for the 2022 Presidential election.
[13]https://g1.globo.com/bemestar /coronavirus/noticia/2020/04/01/espanha-tem-novo-pico-de-mortes-por-coronavirus-em-um-dia-foram-864-nas-ultimas-24-horas.ghtml
[14]https://folha.com/c2fxzs3q

Table V. Quarenteners: Most dense aglomerations

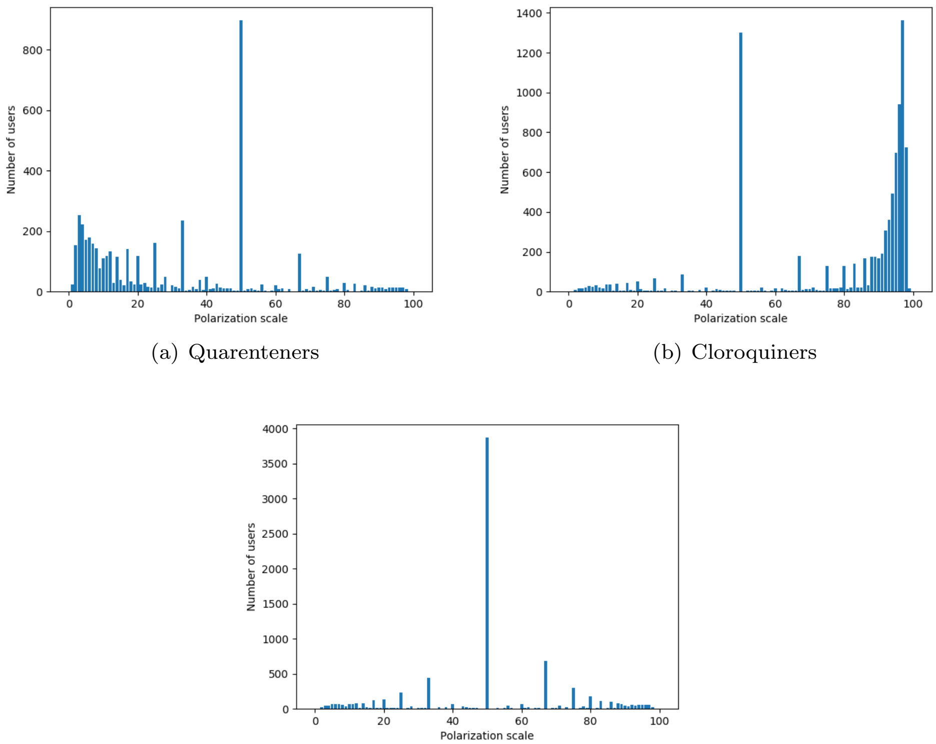
Topic	Agg.	#Clusters	#Tweets	Representative Words	Argument Example
0	Q00	12	657	economy, lives, deaths, die	"You won't starve to stop for 30 or 60 days but the virus can kill thousands in that time. Stop being selfish and a slave to money"
	Q01	3	544	bolsonarists, Trump, hospitals, pandemic	"Even Trump, man. Brazil really likes sewage, so much so that it put a shit in the Planalto Palace"
	Q02	12	466	impeachment, weeks, manifestation, quarentine	"Bolsominions, Have you chosen which relative you will deliver to Coronavirus, to simply please the president? Grandparents? Mother? Mother in law?
1	Q10	4	362	president, worker, transportation, rich	"For fear of the corona, customers and traders will not risk their lives, president"
	Q11	1	360	bozo, scoundrel, cattle, motorcade	"Did you notice that the imbeciles who made the motorcade didn't get out of the car? Why didn't you take a march? Go to the street, bullies!"
	Q12	8	224	home, lives, world, wrong	"Once again the cattle think that the whole world is wrong only the emissary of the good champion of good customs savior of Brazil Bolsonaro is right"
2	Q20	15	349	monster, psychopath, president, dictator	"Bolsonaro plays with the lives of Brazilians, someone needs to stop this man"
	Q21	1	305	we need, together, brazilians, campaign	"Bolsonaro is irresponsible and puts the lives of thousands of Brazilians at risk to ensure a fanciful narrative. DO NOT listen to the president, the life of someone in your family may depend on it!"
	Q22	3	275	stop, health, killer, worse	"Ministry of Health was not consulted on Bolsonaro's criminal campaign against isolation"

rates in hospitals (*hospitals*) and highlight that even Donald Trump (*Trump*) reinforced social distance in the United States[15]. This aggregation encompasses only three (3) clusters, and thus, with highly similar arguments. Finally, the central argument of the third largest agglomeration, Q02, is the consequences of not imposing some level of social distance to combat COVID (*quarentine, weeks*). All these arguments provide further insights on the main reasons why this group opposes Bolsonaro's strategy of vertical isolation as a means to preserve the economy, expressed with a strong political bias against Bolsonaro, his supporters, and other political actors.

Topic 1 has the highest engagement in the Quarenteners group, and it expresses concerns related to the president's actions. The main focus of the largest agglomeration (Q10) are criticisms of Bolsonaro's voters and reasons why the population should not be harmed to keep businesses open (*transportation, worker, rich*). Q11 is the second largest agglomeration, represented by a single cluster of critics of the motorcades supporting the re-opening of businesses (*motorcade, cattle*), an event encouraged by Bolsonaro himself. The central argument of the third largest agglomeration, Q12, voices that Bolsonaro's stake on social isolation is inappropriate and goes against the path taken by most countries in the world (*world, wrong*). This finer-grained analysis reveals that the arguments in this topic highlight that lives should not be sacrificed for the economy, with a strongly negative judgment against President Bolsonaro, his supporters, and their economic interests.

Topic 2 stresses the importance of social isolation. Its biggest agglomeration, Q20, aggregates 15 clusters (25,86% of this topic), which expresses varied criticisms to the president regarding the actions undertaken in response to the pandemic. Derogatory mentions to the president are the most representative words in this aggregate (*monster, psychopath, dictator*). Q21, the second-largest agglomeration, relates to a single cluster that voices the need to fight together against this presidential campaign (*we need, together*). The third agglomeration studied (Q22) argues that health should be prioritized over the economy (*health*) and that social isolation (*stop*) is the proper measure to be taken (contrary to the government's campaign), as endorsed by the health ministry himself. Like all the previous topics,

[15]https://www.nytimes.com/2020/03/29/us/politics/trump-coronavirus-guidelines.html



(a) Quarenteners

(b) Cloroquiners

(c) Neutrals

Fig. 3. Political polarization of users

this analysis of the representative arguments reveals contempt for the president and his campaign.

5.1.3 *Conclusions.* This analysis confirms that the stances of the Cloroquiners and Quarenteners focus on the dilemma of the economy and lives with a strong political bias, with endorsement/rejection to the president. The analysis of the LDA topics using BERTopic revealed that the stance of Cloroquiners is grounded on the one hand on the argument that the virus is not that lethal and that vertical isolation would be a proper solution. They argue that vertical isolation would minimize the impact on the economy and its consequences on the population. From a political perspective, they support the president and his governmental actions and criticize a wide range of opponents, classifying their concerns as excessive hysteria. The Quarenteners, on the other hand, express fear regarding the contagious and its consequences if social isolation measures are not taken. They bring arguments based on the world's experience on COVID control and express deep contempt for the president.

The segmentation of tweets into topics and agglomerations of arguments within topics was beneficial for a deep understanding of the ideas of each group, significantly expanding our original analysis [Ebeling et al. 2020a]. While LDA enabled identifying significant concerns, the BERTopic analysis revealed that arguments in all topics heavily load political bias.

5.2 Q2: Are these groups politically polarized?

Figure 3 shows, for each group, the distribution of users in terms of the polarization metric, whereas the boxplot in Figure 4 details the distribution of the polarization metric for each group in terms of median, first (Q1), and third quartiles (Q3).

Although there are a significant number of politically neutral users in all three groups, Quarenteners

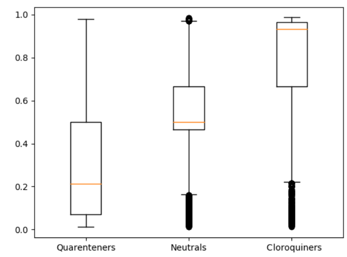


Fig. 4. Distribution of political orientation in groups

and Cloroquiners differ considerably in the number of users oriented towards the left or right, presenting a mirrored behavior. Clearly, the Cloroquiners group is highly polarized towards the right, i.e., users follow a higher proportion of right-wing politicians. Considering the median (92.3), Q1 (66.29), and Q3 (96.21), we observe that 75% of the users in this group are oriented towards the right. Users with a polarization index below 21.5 are treated as outliers.

On the other hand, the Quarenteners group concentrates many users polarized towards the left (i.e., users follow a higher proportion of left-oriented politicians), although it also includes a fair amount of right-oriented users. Compared to Cloroquiners, it has a more heterogeneous ideological character, with the polarization index ranging from 0.6 to 97.3 (Min/Max). Given this range, this distribution does not contain outliers. However, the political orientation of 75% of users (Q3) is oriented towards the left (49.9 or less), and for 50% of users (median), the polarization index is 21% or less.

Finally, the Neutrals group users are evenly distributed towards both right and left, with a median of 50%. The polarization index of 25% of users (Q1) is 46.25 or less, and this value does not exceed 66.3 for 75% of users (Q3). The min/max values are 16.4 and 96.5, respectively, where lower/higher values are considered outliers.

We conclude that the Cloroquiners and Quarenteners groups are politically polarized. The former is skewed towards the right and the later towards the left but slightly less accentuated. We also confirmed the political neutrality of the Neutrals group.

5.3 Q3: Does the polarization affect their social network structure?

5.3.1 *Social Network Metrics and Communities.* Table VI shows properties of the graph constructed for each group, and Table VII summarizes the average metrics of the communities found. These metrics reveal important aspects of the social network structure of each group, and whether their organization in communities reflect a cohesive identity.

The metrics in Table VI reveal that the Cloroquiners have a more connected network structure, considering not only the number of edges, but also the connection between nodes (highest average degree). We also observe that the Cloroquiners network has the highest clustering coefficient, thus resulting is the smallest set of communities (28). The averaged properties of these communities in Table VII confirm they compose larger and more connected communities (highest averages of number of nodes, edges, diameter and average degree). The highest average diameter represents the easier of reaching information within these communities.

In comparison, the communities identified in the Quarenteners network have the highest average density, the lowest average shortest path and the smallest diameter. These metrics reveal that the communities in the group are, in comparison, smaller, with very connected users, but a more limited flow of information within subgroups.

The Neutrals have the lowest clustering coefficient, which explains the largest number of communities found (80). The communities identified in the Neutrals network are very small, with loosely

Table VI. Group Properties

	#Nodes	#Edges	Average Degree	Clustering Coefficient	#Communities
Cloroquiners	1.316.300	3.913.573	12,97	0,01859	28
Quarenteners	1.145.221	8.541.436	6,83	0,01362	44
Neutrals	2.791.367	5.949.930	4,26	0,00004	80

Table VII. Average Properties of Communities per Group

	Avg. #Nodes	Avg. #Edges	Avg. Shortest Path	Avg. Diameter	Avg. Density	Avg. Degree
Cloroquiners	47.010	221.500	1,65	3,25	0,055	3,53
Quarenteners	26.027	69.285	1,54	2,84	0,086	2,42
Neutrals	34.892	56.100	1,93	4,07	0,076	2,23

connected users

We conclude that the Quarentener and Cloroquiners do composed a network structure in which the users are connected with some identity level, and focused on the exchange of information.

5.3.2 *Political Polarized Communities.* We also examined the communities within each group, seeking those related to the politicians from our list. We observed that all three groups have one community that concentrates politicians in an equal right/left proportion. All other communities typically do not involve politicians, with a single exception (Quarenteners). Table VIII highlights the properties of the most polarized communities in each group. These are very representative communities, given the difference of their topological metrics compared to the average metrics in Table VII.

The most polarized community in the Cloroquiners network (22.7% of these users) is also the largest among all polarized communities, with the most connected users and the most extended reach. In the Quarenteners network, we found two polarized communities. The most polarized one (19% of these users) involves significant numbers of left/right politicians. In contrast, the second most polarized one (12.4%) includes only left politicians (28), denoting that users in the Quarenteners group engage politically in distinct ways. The diameter of this second Quarenteners community is the smallest one in all polarized communities. However, it has the highest average degree and smallest average shortest path, making it the most connected community. In other words, it is a small but highly connected community that concentrates on a single political orientation. The average proximity of users in the most polarized community of the Quarenteners and Cloroquiners is similar, but when comparing the diameters, the one related to the Quarenteners is smaller. For the Neutrals, the most polarized community represents 12.2% of this group, with users closer to each other and shorter information reach. The nodes with the highest number of connections in the Quarenteners/Cloroquiners correspond to political opponents in the elections, while in the Neutrals, a newspaper.

Considering the highest betweenness centrality, which is responsible for spreading information throughout the network, we observed in the three most polarized communities of each group a far-right youtuber that propagates ideological content (ProfPaulaMarisa), a far-left politician (ChicoAlencar), and a journalist who discuss an ample spectrum of concerns in a YouTube channel (NilMoreto). Considering the closeness centrality in these three communities, we observe newspapers as the node that most directly influences the polarized networks of Neutrals (G1) and Quarenteners (TheInterceptBR), while for Cloroquiners it is a political user (BolsonaroSP). The second most polarized community of the Quarenteners presents a distinct behavior since the users with the highest betweenness and closeness centrality (mmarescast and joomikhail) are neither newspapers/journalists, politicians, nor influencers. They correspond to ordinary users with a strong political bias, who strongly contribute to a strongly connected community.

5.3.3 *Conclusions.* We conclude that the Cloroquiners and Quarenteners groups are effectively politically polarized, where the former is oriented towards the right, and the later towards the left, but

Table VIII. Polarized Communities Properties

	#Nodes	#Edges	Average Shortest Path	Diam.	Average Degree	Right-Wing Politicians	Left-Wing Politicians	Greater In-Degree	Greater Bet. Centrality	Greater Clos. Centrality
Cloroquiners	299.247 (22,7%)	1443231 (36,8%)	5.35	16	9.64	98	97	jairBolsonaro	ProfPaulaMarisa	BolsonaroSP
Quarenteners	218.066 (19%)	837.360 (9,8%)	5,06	14	7,7	94	71	Haddad_ Fernando	50ChicoAlencar	TheInterceptBr
	141.794 (12,4%)	1.261.800 (14,7%)	2,49	5	17,8	0	28	mmarescast	joomikhail	joomikhail
Neutrals	339.732 (12,2%)	1.233.377 (20,7%)	4,68	12	7,26	95	97	g1	nilmoretto	g1

Table IX. Percentages of LIWC categories for each of the groups

Dimensão	Categ.	Neut.	Chloroq.	Quarent.	Dimensão	Categ.	Neut.	Chloroq.	Quarent.
Group	we	0.06	**0.08**	0.04	**Emotions**	anger	0.16	0.20	**0.23**
sense	assent	0.06	**0.09**	0.06		sadness	0.21	**0.20**	**0.20**
Personal	work	0.34	**0.41**	0.29		anxiety	0.10	**0.12**	**0.12**
Concerns	money	0.25	**0.27**	0.23		neg. emo.	0.39	0.41	**0.44**
	leisure	**0.25**	0.13	0.12		pos. emo.	**0.57**	0.51	0.41
	home	**0.19**	0.06	0.06	**Cognitive**	exclusive	**0.50**	**0.50**	**0.50**
	health	**0.18**	0.14	0.13	**Complex.**	conjunct.	**0.62**	0.54	0.55
	death	0.05	0.07	**0.08**		preps	**0.80**	0.65	0.67
						cog. mech.	**0.92**	0.86	0.86

in a less accentuated and more diverse way. Since the main focus of the Cloroquiners is Bolsonaro's support, the cohesion of this group is stronger. This group is more connected and closed, which reveals a stronger sense of identity. The Quarenteners group presents a more diverse composition, concentrating many users polarized towards left-wing politicians and includes a good number of users oriented to the right. It also has a polarized community only with politicians oriented to the left, more connected and closed, of political activists. Finally, the Neutrals group has a significantly smaller clustering coefficient comparing to the polarized groups. The most polarized community in this group has a lower percentage of nodes and a smaller average shortest path (compared to the polarized groups). The absence of a clear identity decreases the risk perception against it (and therefore, the need to rely on the "echo chamber effect").

5.4 Q4: Do these groups have different psychological aspects?

We investigated the similarities and differences in the use of words that characterize each group's psychological aspects. First, we applied the chi-square statistical test across all 64 LIWC word categories to compare their usage in the three groups. The differences are statistically significant for all LIWC categories, with a single exception (*exclusivity* of the Psychological Processes dimension). Then we analyzed these differences pairwise. Table IX shows the percentages of use of the LIWC categories used to investigate four psychological aspects.

We observed that the most similar groups are the Cloroquiners and Quarenteners, where there are no statistical differences regarding the use of words belonging to 19 LIWC categories. Compared to the Neutrals, all differences are statistically significant, with three exceptions for the Quarenteners/Neutrals pair and two for the Cloroquiners/Neutrals pair. The similarity of the Cloroquiners and Quarenteners in the usage of near 30% of the LIWC categories is evidence that these politically opposed groups share some common psychological aspects. Their strong dissimilarity enforces this evidence with the Neutrals group. These findings are consistent with the Identity Protective Cognition theory since the individuals selectively behave (in this case, verbally) to support their preferred worldview.

- **Cohesion:** The category *we* is significantly more frequent in the Cloroquiners' tweets, an indication of cohesion in this group. Besides, the higher percentage on the use of the *assent* class by the Cloroquiners is statistically significant and enforces the idea that these individuals have a stronger sense of group. There is no significant difference in the usage of *assent* category within the Quarenteners and Neutrals. We can state that Cloroquiners are the group with the highest cohesion and unity.

- **Affect States:** While Quarenteners has the highest percentage of word usage in the *Negative* category and the lowest in the *Positive* category, Neutrals have the opposite behavior (differences of 5 and 15.21 percentage points, respectively). The Cloroquiners follow the same trend as Quarenteners, but in a slightly lower percentage. When comparing Quarenteners and Cloroquiners in terms of negative subcategories, the only statistically significant difference is *anger*. This is evidence that the Neutrals group is the least oblivious to traumatic reports, and the positive emotions are confirmed by the different topics discussed by this group (Section 5.1). For the polarized groups, the higher percentage of negative emotions is aligned with the need to express dramatic events to get their message across or be related to the pessimism regarding their stance (economy vs. life).
- **Cognitive complexity:** Considering the four LIWC categories used to describe this aspect, there is no significant difference in the use of the *exclusivity* category between the three groups. The Neutrals display the other three classes' highest usage (*cognitive mechanisms, conjunctions, and prepositions*). Among Cloroquiners and Quarenteners, only the difference in the use of prepositions is statistically significant. This is evidence that the Neutrals can post tweets with more coherent, complex, and concrete narratives than the two politically polarized groups, which are similar in terms of cognitive complexity. These findings are consistent with [Pennycook et al. 2020], which reports that ideology is not related to beliefs about COVID but to cognitive sophistication.
- **Personal Concerns:** the use of classes in this LIWC dimension confirmed the topics found in Section 5.1: *work* and *money* have the highest percentages among Cloroquiners; *leisure*, *health* and *home* in Neutrals. *death* is more linked to Quarenteners, although in percentages very close to the other groups.

The analysis of the four propositions of the psychological aspects surveyed shows that there are differences between groups. Polarized groups differ in terms of personal concerns and group cohesion, but they are closer when compared to Neutrals in aspects involving Emotions and Cognitive Complexity. The negativity provides evidence that the defense of their views stems from discontent, reflecting the recognition of contradictory thoughts to their identities as harmful. Also, it is possible to confirm that low cognitive sophistication influences the pandemic's perception more than the political orientation. This is consistent with other studies (e.g., [Pennycook et al. 2020]).

6. CONCLUSION AND FUTURE WORK

In this article, we characterized three distinct movements on Twitter related to COVID. For this analysis, we proposed a framework that unifies different dimensions of this characterization. In addition to modeling topics in two granularity levels, we proposed a metric to measure political polarization, the analysis of the social structure using topological network metrics, and the characterization of psychological aspects based on language use. This same analysis framework can be used to investigate the implications of political positioning related to different topics, such as education, environment, and vaccination.

We provided strong evidence that the stances of each group are grounder on political polarization. The most polarized and cohesive group is the Clorquiners, strongly oriented towards the right. Quarenteners are polarized towards the left, presenting a different political connection structure on the social network. The themes that differentiate them reflect the polarization in support or rejection of the president in establishing the dilemma between lives and the economy. Both are similar in terms of cognitive sophistication and negative emotions, presenting evidence that their point of view stems from discontent, and that low cognitive sophistication is more influential in the perception of the pandemic than political orientation.

We intend to evolve the analysis framework and its use in other scenarios. The dimensions of analysis to be assessed include detailing the behavior of the most polarized communities and their difference in relation to users in the same group; expansion of the psychological aspects studied and their relationship with the linguistic categories; adoption of BERTimbau [Souza et al. 2020] as language representation model for BERTopic; investigation of the predictive psychological aspects of polarization; new metrics to calculate a user's political polarization; among others.

REFERENCES

Ajzenman, N., Cavalcanti, T., and Da Mata, D. More than Words: Leaders' Speech and Risky Behavior During a Pandemic. Cambridge Working Papers in Economics 2034, Faculty of Economics, University of Cambridge. Apr., 2020.

AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C. Topic significance ranking of lda generative models. In *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 67–82, 2009.

Amora, P. R. P., Teixeira, E. M., Lima, M. I. V., Amaral, G. M., Cardozo, J. R. A., and Machado, J. C. An analysis of machine learning techniques to prioritize customer service through social networks. *Journal of Information and Data Management* 9 (2): 135–146, 2018.

Angelov, D. Top2vec: Distributed representations of topics, 2020.

Baccianella, S., Esuli, A., and Sebastiani, F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the Intl. Conf. on Language Resources and Evaluation (LREC)*. Vol. 10. pp. 2200–2204, 2010.

Barrios, J. M. and Hochberg, Y. Risk perception through the lens of politics in the time of the covid-19 pandemic. Working Paper 27008, National Bureau of Economic Research. April, 2020.

Bazzan, A. L. C. I will be there for you: clique, character centrality, and community detection in friends. *Computational and Applied Mathematics* 39 (3), Jun, 2020.

Bedi, P. and Sharma, C. Community detection in social networks. *WIREs Data Mining and Knowledge Discovery* 6 (3): 115–135, 2016.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. vol. 3, pp. 993–1022, Mar., 2003.

Blondel, V., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment* vol. 2008, 04, 2008.

Bruine de Bruin, W., Saw, H.-W., and Goldman, D. P. Political polarization in us residents' covid-19 risk perceptions, policy preferences, and protective behaviors. *Journal of Risk and Uncertainty* 61 (2): 177 – 194, 2020.

Conover, M. D., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., and Flammini, A. Political polarization on twitter. In *Proc. of the Fifth International Conference on Weblogs and Social Media*, L. A. Adamic, R. Baeza-Yates, and S. Counts (Eds.). The AAAI Press, 2011.

Costa, L. d. F., Rodrigues, F. A., Travieso, G., and Villas Boas, P. R. Characterization of complex networks: A survey of measurements. *Advances in Physics* 56 (1): 167–242, Jan, 2007.

De Choudhury, M., Jhaver, S., Sugar, B., and Weber, I. Social media participation in an activist movement for racial equality. In *Proc. of the 10th Intl. Conf. on Web and Social Media (ICWSM)*. pp. 92–101, 2016.

Demszky, D., Garg, N., Voigt, R., Zou, J., Shapiro, J., Gentzkow, M., and Jurafsky, D. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Ling.: Human Language Technologies*. pp. 2970–3005, 2019.

Denny, M. J. and Spirling, A. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* 26 (2): 168–189, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Douven, I. and Meijs, W. Measuring coherence. *Synthese* 156 (3): 405–425, 2007.

Ebeling, R., Sáenz, C., Nobre, J., and Becker, K. Quarenteners vs. cloroquiners: a framework to analyze the effect of political polarization on social distance stances. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 89–96, 2020a.

Ebeling, R., Sáenz, C. A. C., Nobre, J., and Becker, K. Quarenteners vs. chloroquiners: A framework to analyze how political polarization affects the behavior of groups. In *Proc. of the 2020 Conf. on Web Intelligence Conference (WI-IAT)*, 2020b.

ElSherief, M., Belding, E. M., and Nguyen, D. # notokay: Understanding gender-based violence in social media. In *Proc. of the 11th Intl. Conf. on Web and Social Media (ICWSM)*. pp. 52–61, 2017.

Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks* 20 (2): 189–201, 2009.

Fortunato, S. Community detection in graphs. *Physics Reports* 486 (3-5): 75–174, Feb, 2010.

Garimella, V. and Weber, I. A long-term analysis of polarization on twitter. In *Proc. of the 11th Intl. Conf. on Web and Social Media (ICWSM).* pp. 528–531, 2017.

Grootendorst, M. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics., 2020.

Grossman, G., Kim, S., Rexer, J. M., and Thirumurthy, H. Political partisanship influences behavioral responses to governors' recommendations for covid-19 prevention in the united states. *Proceedings of the National Academy of Sciences* 117 (39): 24144–24153, 2020.

Hansen, D. L., Shneiderman, B., Smith, M. A., and Himelboim, I. Chapter 3 - social network analysis: Measuring, mapping, and modeling collections of connections. In *Analyzing Social Media Networks with NodeXL (Second Edition)*, Second Edition ed., D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim (Eds.). Morgan Kaufmann, pp. 31 – 51, 2020.

Harb, J., Ebeling, R., and Becker, K. Exploring deep learning for the analysis of emotional reactions to terrorist events on twitter. *Journal of Information and Data Management* 10 (2): 97–115, 2019.

Harb, J. G. D., Ebeling, R., and Becker, K. A framework to analyze the emotional reactions to mass violent events on twitter and influential factors. *Information Processing & Management* 57 (6): 102372, 2020.

Hong, S. and Kim, S. H. Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly* 33 (4): 777 – 782, 2016.

Hornik, K. and Grün, B. topicmodels: An r package for fitting topic models. *Journal of statistical software* 40 (13): 1–30, 2011.

Jerônimo, C., Campelo, C., and de S. Baptista, C. Using open data to analyze urban mobility from social networks. *Journal of Information and Data Management* 8 (1): 83–99, 2017.

Jiang, J., Chen, E., Yan, S., Lerman, K., and Ferrara, E. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies* 2 (3): 200–211, 2020.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32.* ICML'14. JMLR.org, pp. II–1188–II–1196, 2014.

Makridis, C. and Rothwell, J. T. The real cost of political polarization: Evidence from the covid-19 pandemic. *Covid Economics* (34): 50–87, July, 2020.

Manning, C., Raghavan, P., and Schütze, H. Introduction to information retrieval. *Natural Language Engineering* 16 (1): 100–103, 2010.

McInnes, L. and Healy, J. Accelerated hierarchical density based clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on.* IEEE, pp. 33–42, 2017.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun (Eds.), 2013.

Milosh, M., Painter, M., Van Dijcke, D., and Wright, A. Unmasking partisanship: How polarization influences public responses to collective risk. *SSRN Electronic Journal*, 01, 2020.

Mohammad, S. M. and Turney, P. D. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29 (3): 436–465, 2013.

Narayan, A., Berger, B., and Cho, H. Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *bioRxiv*, 2020.

Ordun, C., Purushotham, S., and Raff, E. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. arxiv:2005.03082, 2020.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. inguistic inquiry and word count: Liwc 2001, 2001.

Pennycook, G., McPhetres, J., Bago, B., and Rand, D. Predictors of attitudes and misperceptions about covid-19 in canada, the UK, and the USA,

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., and Ungar, L. Beyond binary labels: Political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Vancouver, Canada, pp. 729–740, 2017.

Puerari, I., Duarte, D., Bianco, G. D., and Lima, J. F. Exploratory analysis of electronic health records using topic modeling. *Journal of Information and Data Management* 11 (2): 131–147, 2020.

Rao, A., Morstatter, F., Hu, M., Chen, E., Burghardt, K., Ferrara, E., and Lerman, K. Political partisanship and anti-science attitudes in online discussions about covid-19. *CoRR* vol. abs/2011.08498, 2020.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining.* ACM, pp. 399–408, 2015.

Sha, H., Hasan, M. A., Mohler, G. O., and Brantingham, P. J. Dynamic topic modeling of the covid-19 twitter narrative among u.s. governors and cabinet executives. *CoRR* vol. abs/2004.11692, 2020.

Slatcher, R., Chung, C., Pennebaker, J., and Stone, L. Winning words: Individual differences in linguistic style among u.s. presidential and vice presidential candidates. *Journal of Research in Personality* 41 (1): 63 – 75, 2007.

Soares, F. B., Recuero, R., Volcan, T., Fagundes, G., and Sodré, G. Research note: Bolsonaro's firehose: How covid-19 disinformation on whatsapp was used to fight a government political crisis in brazil. *Harvard Kennedy School (HKS) Misinformation Review*, 2021.

Souza, F., Nogueira, R., and Lotufo, R. 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)*, 2020.

Tausczik, Y. R. and Pennebaker, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology* 29 (1): 24–54, 2010.

Vargas-Calderón, V., Dominguez, M. S., Parra-A., N., Vinck-Posada, H., and Camargo, J. E. Using machine learning and information visualisation for discovering latent topics in twitter news. *CoRR* vol. abs/1910.09114, 2019.

Walter, R. and Becker, K. Caracterização e comparação de campanhas promovendo o outubro rosa e o novembro azul no twitter. In *XXXIII Simpósio Brasileiro de Banco de Dados: Demos e WTDBD, SBBD 2018 Companion, Rio de Janeiro, RJ, Brazil, August 25-26, 2018*, M. Holanda and J. M. Monteiro (Eds.). SBC, pp. 81–87, 2018.