

NOVA

IMS

Information
Management
School

MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Analyzing Polarization on Social Media

A Case Study of the 2022 Brazil Presidential Election

Felipe Soares Costa

Project Work

presented as a partial requirement for obtaining the Master's Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ANALYZING POLARIZATION ON SOCIAL MEDIA

by

Felipe Soares Costa

Project Work presented as a partial requirement for obtaining the Master's Degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: Roberto André Pereira Henriques

February 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Felipe Soares Costa

Lisbon, February 28, 2023

DEDICATION

Escrevo essa dedicatória em português, minha língua materna, para que seja compreendida por todos os que fazem parte dessa conquista.

Como sempre, a Ti, meu Deus e Criador, quem me sustenta, me protege e me guia todos os dias mesmo eu não merecendo.

A minha mãe que sempre tornou isso possível com seu trabalho duro, garra, cuidado e amor para comigo e meus irmãos. Sempre nos coloca em primeiro lugar, muitas vezes esquecendo de si mesma. Desejo que possamos retribuir todo o seu cuidado um dia!

Aos meus irmãos, Cíntia e Cleiton, que também contribuíram para as minhas conquistas. As oportunidades que eu tive são também frutos do seu trabalho e cuidado.

Aos meus amigos que tornam os momentos todos os momentos mais leves. Aos que estão no Brasil pelos momentos felizes que passamos e pelos reencontros anuais cheios de nostalgia e aos que estão em Portugal por ajudar aliviar por ajudar a aliviar a saudade de casa.

Aos colegas da Nova IMS por todos momentos em passamos juntos a trabalhar incansavelmente para entregar os projetos ao mesmo tempo em que assistíamos as aulas e trabalhávamos.

A todos meu muito obrigado!

Com carinho, Felipe.

ABSTRACT

Social Media has become a big part of our society and has now a significant role in the relationships between inter and intra-communities.

Twitter is now an important communication platform for political campaigns: in the last years, politicians, campaigners, and general users have been extensively using Twitter to promote campaigns and engage in political discussions.

Some studies argue that social media can create filter bubbles by limiting the flow of online information, and therefore creating communities where exposure to political diversity is rare. This selective exposure can build echo chambers where individuals only interact with those who have the same opinions as they have and by doing that, they build a polarized community.

Identifying, understanding, and mitigating polarization is very important for the democratic process. People should be exposed to different ideas and opinions so they can choose their representatives without being influenced by some portion of the information.

This project analyzed political polarization on social media using data from Twitter. Brazil's presidential election in 2022 was used as a case study. Tweets from the two main candidates were extracted. A Topic Modeling algorithm was used to cluster tweets in topics. An Engagement Graph was built based on the interactions between users, candidates, and topics and was used to compute the Topic Centrality measures. A pre-trained Sentiment Analysis model was used to measure the sentiment polarity of each tweet.

In the end, the project analyzed the extracted features and identified which topics were more central to each candidate and how users interact with them. The major conclusion of this work is that polarization in Brazil is more affective than ideological since the user's sentiments towards topics are not as relevant as the sentiments towards the candidates.

KEYWORDS

Political Polarization; Sentiment Analysis; Natural Language Processing; Topic Modeling;

INDEX

1. Introduction.....	1
2. Literature review	3
2.1. Polarization in Social Media	3
2.2. Topic Modeling	7
2.3. User’s Engagement in Social Media	8
3. Case Study: Political Polarization in brazil	10
4. Methodology	11
4.1. Tools and Technology.....	12
4.2. Data Extraction	14
4.3. Topic Modeling.....	16
4.3.1. Preprocessing	16
4.3.2. Model Selection.....	18
4.4. Engagement Graph.....	23
4.5. Sentiment Analysis	25
5. Results and discussion	27
5.1. Topic Modeling	28
5.2. Topic Centrality.....	30
5.3. Sentiment Analysis	33
6. Conclusion	37
7. Limitations and recommendations for future works	38
8. References	39
9. Appendix (optional)	43
9.1. Snsrape Commands	43
9.2. Spark NLP Pipeline.....	43
9.3. Representative documents per topic.....	44
9.4. Network Analysis.....	44

LIST OF FIGURES

Figure 1 - Retweets and Mentions Network (Conover M. et al.)	5
Figure 2 - Affective Polarization Metric	6
Figure 3 - Methodology Overview	11
Figure 4 - Google Cloud Storage Bucket	13
Figure 5 - Architecture Overview	13
Figure 6 - Tweets from the Candidates	14
Figure 7 - LDA Model Operation (Buenano-Fernandez D. et al.)	18
Figure 8 - Coherence Score x Number of Topics	19
Figure 9 - Word Clouds for LDA Topic Model	19
Figure 10 - Word Clouds for GSDMM Topic Model	20
Figure 11 - Similar Topics that can be merged	21
Figure 12 - Coherence Score for BERTopic	22
Figure 13 - Word Clouds for BERTopic	22
Figure 14 - Engagement Graph Representation	23
Figure 15 - Engagement Graph	24
Figure 16 - Word Clouds Hashtags	26
Figure 17 - Distribution of users per Community	27
Figure 18 - Distribution of Topics Per Candidate	29
Figure 19 - Number of Replies Per Topic	29
Figure 20 - Topic Centrality per Community	30
Figure 21 - Total Network's Topic Centrality x Bolsonaro's Topic Centrality	31
Figure 22 - Total Network's Topic Centrality x Lula's Topic Centrality	31
Figure 23 - Lula's Topic Centrality x Bolsonaro's Topic Centrality	31
Figure 24 - Number of Replies per Topic and Community	32
Figure 25 - Sentiments per Topic and Community	33
Figure 26 - Sentiments per Communities and Candidates	33
Figure 27 - Distribution of Hashtags per Group	35

LIST OF TABLES

Table 1 - Dataset Schema	15
Table 2 - Top 10 replies by Location	15
Table 3 - Sentiments Polarity on Tweets	25
Table 4 - Distribution of Hashtags Per Group	34
Table 5 - Top 5 Hashtags per Community	35

LIST OF ABBREVIATIONS AND ACRONYMS

GCS	Google Cloud Storage
POS	Part of Speech
CLI	Command Line Interface
TM	Topic Modeling
LDA	Latent Dirichlet Allocation
GSDMM	Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture Model
MGP	Movie Group Process
NLP	Natural Language Process
PT	Partido dos Trabalhadores (Worker's Party)
LeIA	Léxico para Inferência Adaptada

1. INTRODUCTION

Social Media has become a big part of our society and has now a significant role in the relationships between inter and intra-communities. People from different social and cultural backgrounds can interact, share their opinions, engage in discussions, and advocate for their beliefs. Twitter is now an important communication platform for political campaigns: in the last years, politicians, campaigners, and general users have been extensively using Twitter to promote campaigns and engage in political discussions. This was observed in the 2018's presidential election in Brazil, which marked an increase in digital campaigning tactics. For the first time in history, due to a law that authorized paid political advertisement on social media, candidates were able to pay to sponsor posts, to use bots and other digital techniques to increase their online visibility [1]. The election was also marked by a big dispute between the left and right wings and very polarized discussions over the internet.

Twitter posts, also called tweets, are disseminated publicly, and can easily be viewed by all the users creating a huge network. This network creates a live stream of ideas since users who interact with a candidate by replying to his posts are propagating the message through his network. Park C. S. [2], investigated how Twitter can act as a motivator of political engagement and the role of opinion leadership as a driver to motivate political activities on Twitter.

Several studies have been conducted regarding the influence of social media platforms in political discussions. Some studies argue that social media can create filter bubbles by limiting the flow of online information, and therefore creating communities where exposure to political diversity is rare [3]. This selective exposure can build echo chambers where individuals only interact with those who have the same opinions as they have and by doing that, they build a polarized community. On the other hand, some authors challenged this affirmation and studied how social media can help to reduce political polarization by increasing the incidental exposure to political messages shared by peers [4].

Identifying, understanding, and mitigating polarization is very important for the democratic process. People should be exposed to different ideas and opinions so they can choose their representatives without being influenced by some portion of the information. In the same way, users should be encouraged to discuss different points of view to challenge their own opinions about topics that are discussed by their leaders.

Apart from having many studies conducted about political polarization with social media, most of them focus on English posts, especially from the United States, rather than other countries and languages. Also, most of the researchers are focusing on measuring polarization rather than understanding which factors can contribute to a polarized state. Also, the authors use polarization as a broad concept instead of separating between affective and ideological polarization [5].

This project aims to analyze political polarization on social media based on the perspective of user engagement and the measures of Topic Centrality proposed by Raymond C. [6]. Brazil's presidential election in 2022 will be used as a case study. Tweets from the two main candidates, and party leaders, will be extracted to build the input dataset. A Topic Modeling algorithm will be used to cluster the tweets in topics. An Engagement Graph will be built based on the interactions between users, candidates, and topics and will be used to compute the Topic Centrality measures. A pre-trained Sentiment Analysis model will be used to measure the sentiment polarity of each tweet.

These features will be used to identify which topics are more central to the party leader base, which topics promote more engagement, and how the feelings expressed by the candidate can impact the users' engagement.

Regarding the technology and resources, different tools and libraries were used in this project:

- Python Language Programming
- Spark on Databricks for Data Processing [7]
- Google Cloud Storage (GCS) for Data Storage [8]

Tweets will be extracted using Snsrape [9], an open-source service to extract data from social media platforms. The major contributions of this project are:

- A methodology to analyze political polarization on social media
- A complete analysis of the Polarization in Brazil's presidential election in 2022

Additionally, is expected to demonstrate which topics are more central to each candidate and how general users respond to the content.

This project is organized as follows. Section 2 presents a literature review of studies about Polarization and User Engagement on Social Media and the technical aspects: of algorithms and techniques covered in this project. Section 3 presents the Case Study of Political Polarization in Brazil. Section 4 will introduce the methodology and steps applied to extract, process, and analyze the data. Section 5 presents and analyzes the results. Section 6 will add the conclusions and findings of this work. Finally, Section 7 will discuss the limitations encountered during this project and will recommend some ideas for future work.

2. LITERATURE REVIEW

This section provides a literature review that serves as a theoretical basis for this project. It starts with an introduction to Polarization in Social Media, including the definition and methodologies to measure and interpret polarization. The next section presents some studies related to analyzing polarization using Topic Modeling algorithms. Users' Engagement in social media and how polarization can be analyzed from this perspective will be discussed. Following, an overview of important data preprocessing techniques used in this study will be presented.

2.1. POLARIZATION IN SOCIAL MEDIA

To study polarization on social media, it is important to establish how to identify, measure and describe an environment's polarization level. BRAMSON A. et al. [10] argue that polarization is not a single unambiguous concept, but a collection of social dynamics and configurations combined that can be measured and described separately. Therefore, a variety of non-competitor models can coexist since they are explaining different notions of polarization. To explain these different notions the study defined the nine senses of polarization and proposed formal measures for each one, to refine the methodology used to describe polarization. By providing a clear method of capturing polarization the study aims to facilitate the evaluation of polarization models in their ability to clarify relevant social dynamics. The proposed measures are based on properties that can be anything that admits cardinal values because its formalism only depends on the features of the distribution and not on what the distribution represents. The author classified the measures into two groups given the subset of data on which the measure should be applied: Population Measures and Group Measures. These are the Population Measures:

- Spread: It's represented by the difference between the agent with the highest belief value minus the one with the lowest belief value. The wider the difference in the most extreme views held, the more polarized the population's ideas are. Spread only applies to the entire population rather than a specific group.
- Dispersion: It can be measured using any variation of statistical dispersion: mean difference, average absolute deviation, standard deviation, coefficient of variation, or entropy. Dispersion takes into consideration the shape of the whole distribution as opposed to spreading. Dispersion can increase or decrease without affecting spread.
- Coverage: A polarized society is often described as one with little diversity of opinion. In that sense, polarized beliefs can be represented as narrow bands in the opinion space. To measure polarization in the sense of Coverage it is necessary to divide the spectrum of beliefs into small bins. The proportion of empty bins will then be the coverage value.
- Regionalization: Still considering the opinion space as a set of small bins, polarization can be measured in a sense of Regionalization by counting the number of empty spaces between each filled area. This will correspond to the area of uncovered opinion spaces.

To compute the remaining measures is necessary to define groups. Groups can be defined by looking at peaks on a histogram representing the opinion beliefs or by network links representing the association, influence, or communication.

- Community Fragmentation: The degree to which the population can be broken into subpopulations. As a conceptually independent sense of polarization, the more groups there are, the greater the polarization.
- Distinctness: Distinctness is defined by the degree to which each group can be separated or the degree of separation between sub-distributions inside a group.
- Group Divergence: Divergence is the distance between two subgroups' beliefs. It can be measured by the difference between the mean of the two groups.
- Group Consensus: Measure how equal the opinions of individuals inside a group are. A feasible way to measure Group Consensus is the absolute deviation within each group, aggregated over all the groups.
- Size Party: It can be easily measured by comparing the number of individuals in a group. If there are too many groups it is preferable to consider measures that aggregate more intuitively, for example, the sum of the differences in the size of each group from the mean group size.

Finally, the study analyzed data regarding people's political views and views towards abortion from the General Social Survey (GSS). The article concludes that there are few clear trends for any segment of the data, and no strong trends of increasing or decreasing polarization for political views or abortion from 1984 to 2012.

To create a polarization model is also important to define which features, algorithms and techniques are suitable to solve the problem. MAROZZO F. and BESSI A. [11] studied the behavior of social network users and how news sites are used during political campaigns characterized by the rivalry of different factions. The researchers proposed a methodology to discover this behavior and apply it in a case study: the constitutional referendum that was held in Italy on 4th December 2016 where voters were asked if they approve a law that amends the Italian Constitution to reform the composition of the Italian Parliament. The study aimed to analyze how users express their voting intentions about the referendum in the weeks it took place. Also, to understand how voting intentions evolve before the vote and if they change over time.

Three sets of keywords were defined based on the voting intentions: neutral, yes, and no. Using these keywords, the authors collected 338,592 tweets (1,165,176 considering the retweets) based on posts about the referendum, in the weeks before the voting day, containing the defined keywords (hashtags).

In the end, the study was able to identify that 48% of Twitter users were polarized towards no, 25% towards yes, and 27% had neutral behavior. It was also identified that most users categorized as "no" supporters have never changed their opinion during the weeks preceding the vote. On the other hand, a significant part of neutral users moved towards no. By analyzing the URL to news related to the referendum, the study was able to identify sites that were polarized during either one of the options. The study proposed a well-defined methodology to measure polarization. However, this methodology requires a manual definition of the keywords before collecting the posts. Also, it does not consider users that post something about the referendum without using the defined hashtags.

BELCASTRO L. et al. [12] proposed a methodology, called IOM-NN (Iterative Opinion Mining using Neural Networks) (Iterative Opinion Mining using Neural Networks), for discovering the polarization of social media users. The methodology consists of an automatic incremental procedure based on a feed-forward neural network. The process starts by collecting Twitter posts based on a set of predefined keywords in favor of specific factions. The posts are then classified using the neural network. Finally, the classified posts are analyzed for determining the polarization of users towards a faction. The methodology was applied in two case studies: the 2018 Italian general election and the

2016 US presidential election. The methodology results were compared with the election results and had much better accuracy than the average of the opinion polls. Apart from being highly scalable and complete, this approach also requires a manual definition of partisan keywords and does not include posts without these keywords.

CONOVER M. et al. [13], studied political polarization on Twitter by examining data from the 2010 U.S. midterm elections. They collected 250,000 politically relevant tweets produced by more than 45,000 users. Using this data, they build two networks:

- Retweet network: connect users that retweeted content from others
- Mention Network: connect users who mentioned others, including replies

After creating the network, a community detection algorithm (label propagation) was used to detect clusters in the networks. Analyzing the shape of the produced networks they realized that the Retweet Network's shape contains two distinct communities while the Mention Network's shape presents only one single community. In Figure 1 it is possible to observe the two networks: Retweet Network on the left and Mentions Network on the right.

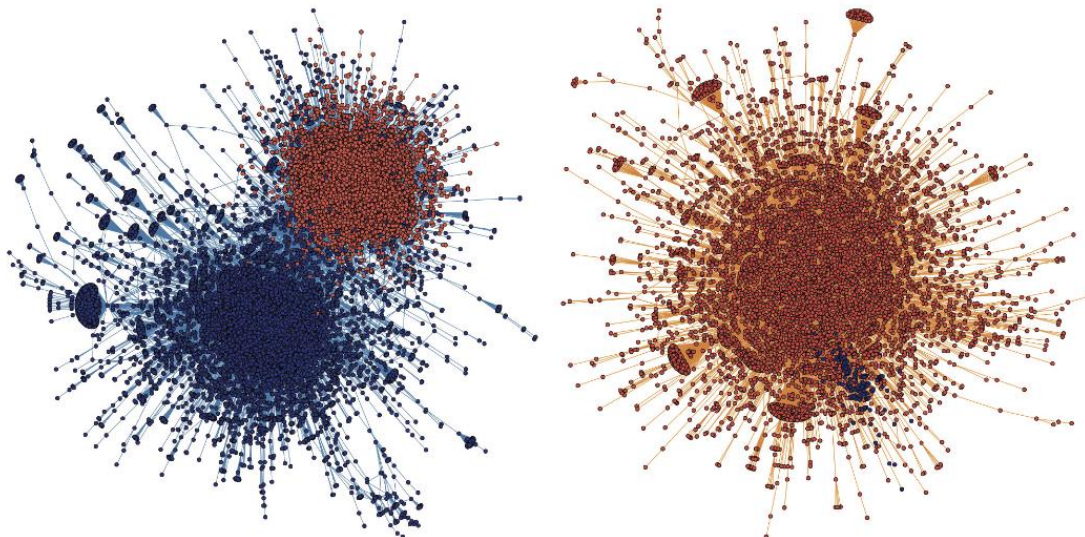


Figure 1 - Retweets and Mentions Network (Conover M. et al.)

This output indicates that users are more likely to retweet messages that came from users in the same ideological spectrum. On the other hand, the Mention Network does not present any sign of segregation and indicates that users are exposed to any available information, even from opposing parties.

In order to explore the content injection behavior, the authors introduced the content of political valence in hashtags. Political valence is a measure that indicates the partisanship of a hashtag based on the proportion of users from a given party that uses this hashtag. The study found that users who use more neutral hashtags are more likely to receive connections from users on both sides.

Finally, quantitatively speaking the study concluded that messages on social networks remain highly partisan, apart from understanding that some users are still willing to share information and communicate with users from opposing sides.

TYAGI A. et al. [14], introduced affective polarization as a form of polarization driven by highly negative sentiments from users towards opposing groups. The study focuses on analyzing the polarization in the climate change discourse. Users were separated into two distinct groups: Believers, the ones who accept anthropogenic causes of climate change, and Disbelievers, the ones who reject the same.

The research proposed a framework to analyze affective polarization in online climate change discourse. They collected 38M of tweets between August 2017 and September 2019, containing the keywords and hashtags: "Climate Change", "#ActOnClimate", and "#ClimateChange". Data were aggregated by week totaling 100 weeks (about 2 years).

Using a set of seed hashtags to label users as Believers and Disbelievers, a classifier model was trained. The hashtags ClimateChangelsReal and SavetheEarth were assigned as Believers and ClimateHoax and Qanon as Disbelievers. This process produced seed users that were later used in conjunction with a text classifier to generate the final labels. Of 7M of tweets 3.9M were classified as Believers and 3.1M as Disbelievers.

Finally, using sentiment analysis the sentiment of each group towards the climate change discourse and toward the opposing side was computed. The results demonstrate that Disbelievers are more hostile toward Believers in the climate change discourse. In Figure 2 it is possible to observe this behavior where high values indicate the intensity of the negative sentiments towards the opposing group. Disbelievers also used more words and hashtags related to natural disasters than Believers.

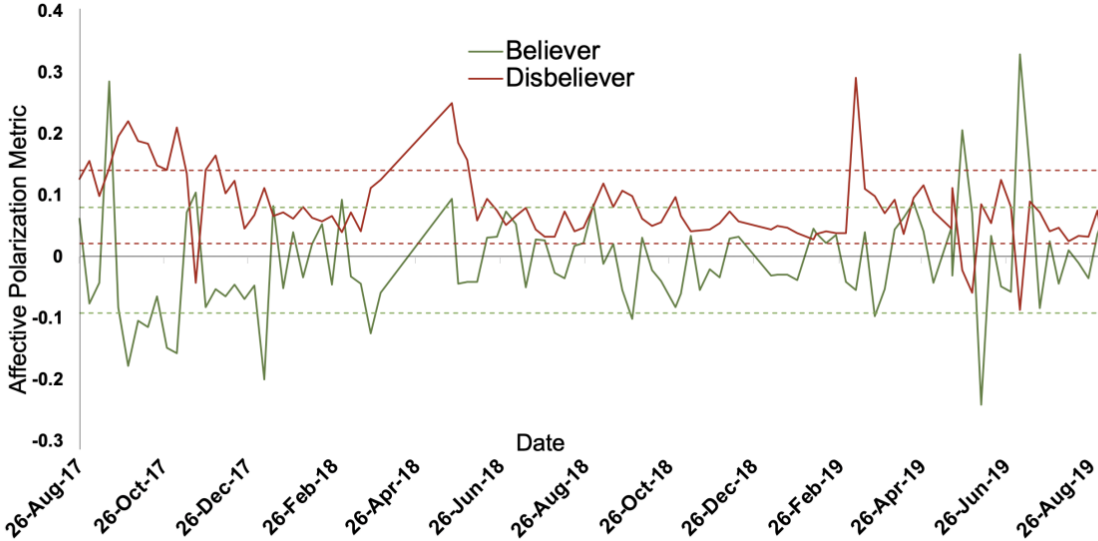


Figure 2 - Affective Polarization Metric

2.2. TOPIC MODELING

Topic Modeling (TM) is an unsupervised learning technique that finds groups (or clusters) in text. Topic Modeling is a branch of Text Mining, the procedure of extracting information from Text. Topic Modeling has been used in literature to explore and understand discussions in social media.

HE Z. et al. [15], investigate polarized topics in social media. The study proposed a method to automatically detect polarization among partisan (liberal vs. conservative) media sources called Partisanship-aware Contextualized Topic Embeddings (PaCTE). The proposed methodology was tested in the AYLIEN COVID-19 dataset³, consisting of 1.5M news articles about the global pandemic from 2019 and 2020. The authors divided the publisher between liberal and conservative sources. They selected six well-known sources from both parties and after filtering the initial dataset, 66,368 articles were left spanning from Jan 2020 to July 2020. Following the methodology, an LDA Topic Modeling was trained, and 30 topics were selected. Using a fine-tuned pre-trained Bert-base-uncased model, the new articles were classified according to their political bias. For evaluation, the model was compared with the Leave out estimator, which is purely based on statistical features. The new method was proven to be more precise and accurate in capturing topical polarization.

KIM S. and Cho H. [14] focused on analyzing user behavior in online communities in terms of topics. The research used the LDA Topic Modeling technique, with an important substitution: replacing the concept of “words” in the textual data with “users” in the online community. The general concept behind the methodology is that users post comments on social media according to their behavioral characteristics and interests, in the same way, that words appear in documents according to their latent topics. So, instead of representing comments on a post topic as a list of words, the methodology represents it as a list of users who interact with this post. The research was able to identify that users’ behavior and interests can reflect their interactions, especially in common engagement with related articles. The new user topic can reflect dynamic features (that can change over time) better than the traditional topic model. The user-topic model can be used to cluster users, with their behavioral features, according to their interests for example. Although this method considers the topics and the user engagement with them, it doesn’t assess how the users feel about it.

Topic Modeling was also used by NASKAR D. et al. [17]. This study described a methodology for analyzing sentiments in social media using Topic Modeling. The authors used two datasets in the study with data collected from the 2010 US Congressional elections:

- A Twitter dataset containing 60,00 tweets from 6,000 users.
- A Facebook dataset containing public posts and comments on the Facebook pages of two major parties with 17,000 entries.

LDA (Latent Dirichlet Allocation) was used to select topics from tweets. A Sentiment Analysis technique was used to classify the tweet as positive, negative, or neutral. The sentiments were aggregated at a topic level to obtain the sentiment of each topic. The authors use two different aggregation methods: a simple average of sentiment scores for all posts in each topic, and a weighted average that considers the topic probabilities of each post.

The performance of the proposed method was evaluated on both datasets. In the end, the study found that Topic Modeling can significantly improve sentiment analysis compared to traditional approaches based on bag-of-words, however, the accuracy depends highly on the Topic quality.

2.3. USER'S ENGAGEMENT IN SOCIAL MEDIA

Different studies analyzed the behavior of social media users regarding political engagement. LAWRENCE, E. et al. [19], studied the relationship between blog posts and readers' political behavior. The authors conducted interviews and divided the respondents into three categories: those who read exclusively left-leaning blogs, those who read exclusively right-leaning blogs, and those who read both left and right blogs. The results showed that about 94% of political blog readers consume only blogs from one side of the ideological spectrum and the remaining 6% read blogs from both sides. They realized that this apparent homophily should derive from selective exposure, as people choose blogs whose political perspective matches their own and should produce a pattern of political polarization. To verify this assumption the study compared the distribution of three measures of political preferences: party identification, self-reported ideology on the liberal-conservative spectrum, an ideology on the liberal-conservative spectrum, and an ideology scale based on issue positions. They observed that left-wing blog readers are most liberal and Democratic, and right-wing blog readers are conservative and Republican. The study did not find a correlation between exposure to diverse viewpoints and user participation. Instead, they found that left-wing blog readers and cross-cutting readers participate more than readers of right-wing blogs.

RAYMOND C. [6], studied how political messages can affect the discussion and interactions in social media by bonding, i.e., rallying members within a group, while other messages are bridging. The researcher proposed the use of an engagement graph to discuss the behavior and two measures of topic centrality: total network topic centrality, and party leader topic centrality. This methodology was applied in the context of the 2019 Canadian Federal Election, using tweets from the five major English-speaking party leaders: Andrew Scheer, Elizabeth May, Jagmeet Singh, Justin Trudeau, and Maxime Bernier. Using the Twitter API, 7,978 tweets from the party leaders were collected during the year preceding the election. All available retweets from general users were also collected for a total of 113,293 retweets.

After collecting the tweets an engagement graph was built using three types of vertices:

- producers: those who produce content, i.e., tweets.
- the content itself, or the tweet itself
- those who engage (retweet) the content

To evaluate the bridging and bonding characteristics of messages the tweets were then organized into topics. Using LDA topic modeling, the author extracted 7 topics with a coherence score of 0.48. To compute the metrics the author used the eigenvector. The proposed measures are:

- Total Network Topic Centrality: It is computed by aggregating the eigenvector centrality of all tweets of a certain topic in the entire engagement graph.
- Party Leader Topic Centrality: The Party Leader Topic Centrality refers to how central a topic is to a party leader's base. It's computed by taking a subgraph of the party leader with all the users who interact with him. And then aggregating the eigenvector centrality in the same way as Total Network Topic Centrality.

Some discrepancies in engagement were found: party leaders often ignored topics that drove higher levels of engagement among their bases. The study also concluded that topics that could be very important to discuss were less important to the party leader's base and vice versa.

Overall, the proposed measures can be helpful to understand political engagement on social media, however since the author used only the retweets from the users to build the graph, and not the

replies, some context information about how the user feels or what topics promotes more discussion among the party leaders base.

MENTZER K. et. al [19], investigated affective polarization on Twitter. They analyzed data from the 2018 U.S senate election. The study aimed to identify how to measure affective polarization on Twitter and which factors can influence this polarized state. Tweets from September 27 through November 13 of 2018, were collected representing a 7-week time frame. The initial dataset was composed of 17,178,617 tweets but after removing tweets that do not include mentions of the candidates this number decreased to 12,595,639 tweets.

A network composed of user retweets was created. Using a community detection algorithm 10 communities were extracted from this network. From these communities, the top 10 most retweeted candidate accounts were selected to identify the ideological affinity of each community. They also extracted sentiments from tweets and segmented the users by gender.

The results show a greater level of polarization among Conservatives over Liberals. Moreover, the study was able to conclude that both women and men talk more positively about male candidates than they did about female candidates. Conservatives were more likely to talk positively about female candidates. Overall, the study identified that affective polarization in this context was influenced by both party and gender.

SOUTH T. et al [20], studied the information flow in social networks using news on Twitter. The author developed a method for estimating the amount of information exchanged between users.

The dataset was composed of 3.2M tweets containing news articles shared on Twitter between December 2016 and February 2017. To collect this data the authors first identified 1,000 news sources well known for producing reliable news content and then retrieved all tweets containing any mention of these sources during the given period. In the next step, a network was built, and tweets were built. Each node in the network represented a news article and each edge represented a link between two tweets linking both articles. Using a PageRank algorithm, the importance of each article was computed to identify the most influential news sources and the most popular news stories on Twitter during the study period.

Next, a classifier was trained using a set of hand-labeled tweets in a set of categories: Breaking News, News Analysis, Feature Article, Opinion/Editorial, General Information, Entertainment, Sports, Business and Finance, and Technology and Science

Finally, the study analyzed several metrics: the distribution of links between sources and topics, the centrality of sources and topics in the network, and the extent to which news stories tended to spread across different topics and sources.

The results showed that 56.3%, the majority of tweets, were classified as "General Information". "Breaking News" corresponds to 20.7% of the tweets followed by "Opinion/Editorial" with 10.2%. The remaining categories represented less than 5%. Moreover, the study found that some categories drove a higher level of engagement like "Breaking News" and "General Information".

3. CASE STUDY: POLITICAL POLARIZATION IN BRAZIL

Recently Brazil has been suffering from different scandals and political events. Everything started in 2014 with Operation Car Wash, a corruption scandal implicating the entire political class. In the same year Dilma Rousseff, the candidate from PT (Workers Party), won the election becoming the first Brazilian female president. Rousseff soon started to lose the support of her legislative allies for implementing austerity policies different from those she defended during her campaign. In the meantime, the operation car wash exposed several cases of corruption in Brazil and one of the main focuses was the PT leadership. This whole scenario culminated in Rousseff's impeachment in August 2016 [21].

By 2018, the ongoing Operation Car Wash made several arrests including leaders of different parties. However, the most significant case was former president Luiz Inácio Lula da Silva accused of bribery, money laundering, and influence peddling. Lula was convicted by Judge Sergio Moro in the first instance, sentencing him to nine years and six months. His conviction was confirmed in January 2018 by the Federal Regional Court and Lula remained free until April 2018 when the Supreme Court ruled that he must start serving his sentence immediately.

During the presidential election in 2018, Fernando Haddad, former mayor of the city of São Paulo, was PT's choice as the party's candidate. His campaign platform was based on associating himself with Lula, on a failed attempt to win the election by getting the votes of the North and Northeast regions, dedicated supporters of the former president. Given the development of Operation car wash, a strong feeling of hate towards PT was created by that time and Jair Bolsonaro, a congressman for the state of Rio de Janeiro, started a strong opposition to PT and the left wing. Bolsonaro and Haddad made it through the first round of the elections, and they continued in the race for the presidential chair in the second round.

During a campaign event on September 6 of 2018, Bolsonaro was stabbed and lost 40% of his blood. He was hospitalized and had a colostomy bag until January 2019. This event gave him a strong motive for declining participation in traditional campaign activities, including the presidential debates, and boosted his popularity and media appeal.

Jair Bolsonaro won the 2018 election and became the Brazilian new president rising as a leader and a face of the right-wing in Brazil. His mandate was marked by his constant discussions against the media, strong comments about sensitive themes, and the COVID-19 pandemic. Bolsonaro was portrayed as cruel, indifferent, and incompetent in the context of the pandemic, which caused a decline in his approval rates [22]

On June 24 of 2021, the Supreme Court overturned the convictions declaring that Sergio Moro was biased during the investigations and therefore all evidence gathered under Moro's supervision was disregarded. Lula was released from prison and became the leader of the left-wing and candidate for the 2022 presidential election.

Several studies were conducted in the last years about the current political scenario in Brazil [23], [24]. There is a consensus about the polarization between the left and right wings, however, there's no consensus if it's an affective or ideological polarization. Areal J. [25] proposed that polarization in Brazil is neither affective nor ideological, it's driven by negative political identities. He states that Brazilians dislike the opposing side not because of ideological disagreements or identification with their side.

This polarization culminated in another campaign rally in 2022 with the two of them at the center of the debates. Lula won the election with 50,9% of the votes and become president of Brazil for the third time.

4. METHODOLOGY

This section describes the methodology applied to complete this project. Initially, a complete overview of the infrastructure and tools used during the implementation and their role in each stage will be presented. Following this, the data acquisition process is explained, and the dataset is described. In the next steps, three key assets will be created:

- A Topic Modeling algorithm will be used to select topics from candidates’ tweets.
- An Engagement Graph will be built to identify which topics are more central to each candidate network using the Eigenvector Centrality.
- A pre-trained Sentiment Analysis model will be used to identify the sentiment behind each tweet and hashtag.

Finally, the produced assets will be used to analyze the answer to the following questions:

1. Which topics are more central to each candidate’s network?
2. Which topics can bridge or bond communities?
3. Do sentiments expressed in tweets change concerning the topics?
4. Can sentiments expressed in hashtags provide a good understanding of users’ feelings toward candidates?

Figure 3 presents an overview of the methodology described. Each step will be discussed in detail in the next sections.

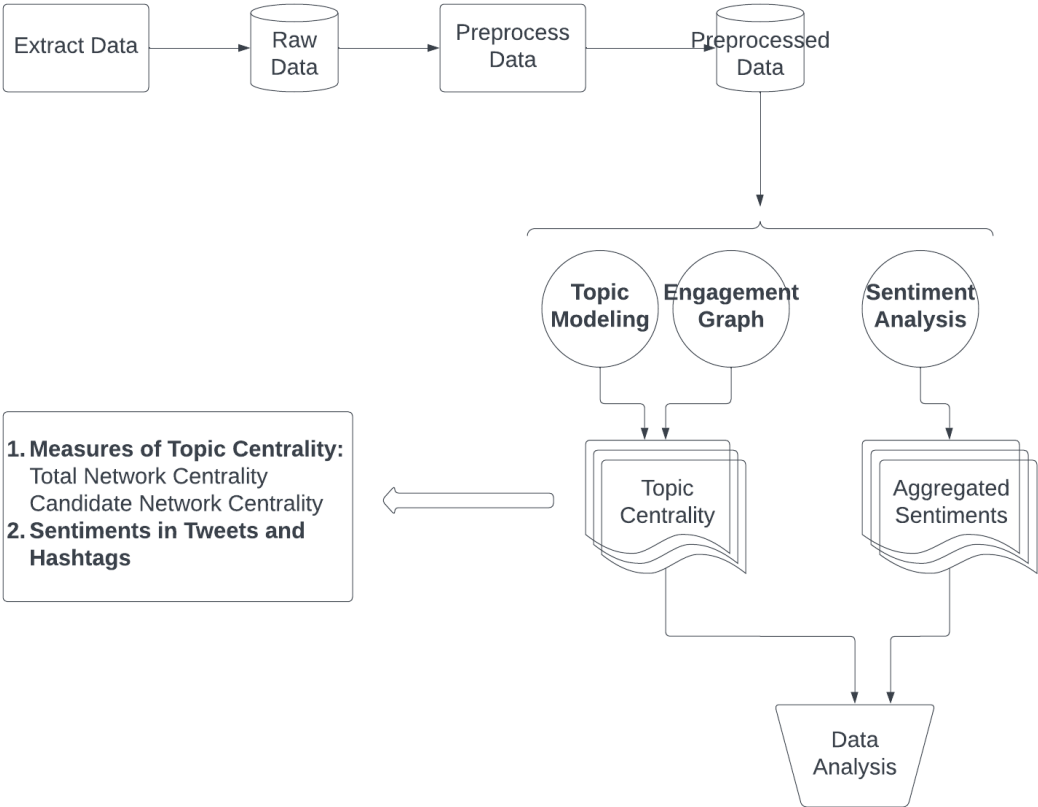


Figure 3 - Methodology Overview

4.1. TOOLS AND TECHNOLOGY

The architecture defined to implement this project is composed of three main components: Extraction, Storage, and Processing. Each one corresponds to a set of tools or technologies with a defined responsibility within the architecture. Extraction corresponds to the services and libraries responsible for extracting tweets. Storage corresponds to the file formats and storage solutions where the data will be placed. Processing is the set of tools used to orchestrate the steps, process the data, and extract all the information needed.

Twitter provides a REST API to extract tweets and users' data from its platform. However, the free tier of this API comes with some limitations related to the historical data available and the number of requests for a given time window. To overcome these limitations Snsrape was used to extract the tweets [9]. Snsrape is a Python service scraper that supports different social media services like:

- Twitter
- Facebook
- Instagram
- Mastodon
- Reddit
- Telegram
- VKontakte
- Weibo

It comes with a python library and a command line interface (CLI), that helps users to extract information from social media providers by applying different sorts of criteria. Snsrape does not have the same limitations as the Twitter API and can extract tweets without limiting historical access and deals internally with the rate limits associated with the number of requests per minute. Depending on the social media service and the entity being extracted Snsrape can produce outputs in formats like JSON and CSV. In this project, given the semi-structured nature of a tweet object, the data were extracted in a JSON format.

For the Storage Layer, a GCS bucket in the Google Cloud Platform was created. GCS is a managed service to store structured, semi-structured, and unstructured data. It is scalable, secure, highly available, provides backup options, and supports highly intensive and big data applications. GCS makes moving data to a cloud environment seamless, effortless, and cost-effective. GCS mimics a File Storage system where root folders are represented as buckets that can hold other folders or file objects [8].

The structure of the GCS bucket can be visualized in Figure 4, and it is composed of the following folders:

- raw: raw json files extracted with SnScrape
- preprocessed: preprocessed tweets
- corpus: input corpus to topic modeling algorithm
- predicted: predicted topics and sentiment analysis

<input type="checkbox"/>	Name	Size	Type
<input type="checkbox"/>	corpus/	—	Folder
<input type="checkbox"/>	predicted/	—	Folder
<input type="checkbox"/>	preprocessed/	—	Folder
<input type="checkbox"/>	raw/	—	Folder

Figure 4 - Google Cloud Storage Bucket

At each step of the implementation, data was recovered from one folder in the bucket, transformed, processed, and then ingested in the next folder. This strategy helps to find errors easily since the development process is split into several layers and helps to avoid unnecessary reprocessing: if an error was found on a given state of the data only the failed steps need to run again and not the whole pipeline.

Due to the huge amount of data produced by Brazil's users, the architecture should be able to process huge chunks of data in an acceptable time. Running all these transformations and data processing tasks in a server will be expensive and time-consuming. Distributing computing technology is the best approach here since it can process massive amounts of data by splitting the process into different workers' machines. Spark is currently the most well-known distributed computing technology to work with Big Data. However, creating, deploying, and managing a Spark cluster is not a trivial task and requires some expertise and it can be expensive depending on the infrastructure. The solution for that is to use a managed Spark environment like Databricks. Databricks is a fully managed open-source distributed computing framework created by the founders of Apache Spark [7]. All the data processing tasks in this project will be performed using Databricks' built-in Python notebooks with several Pyspark transformations [31]. The data will be processed and stored as parquet files in the GCS bucket.

In Databricks a few Python libraries will be installed and used to perform the Data Processing, Topic Modeling, Engagement Graph, and Sentiment Analysis. Databricks is essentially a Spark environment so any Python library can be easily installed and managed at a cluster level. These libraries will be covered in detail in the next sections.

Figure 5 presents an overview of the architecture. It is not in this project's scope a detailed discussion about each tool separately but focus and how they can be used to produce the features and extract the information needed.

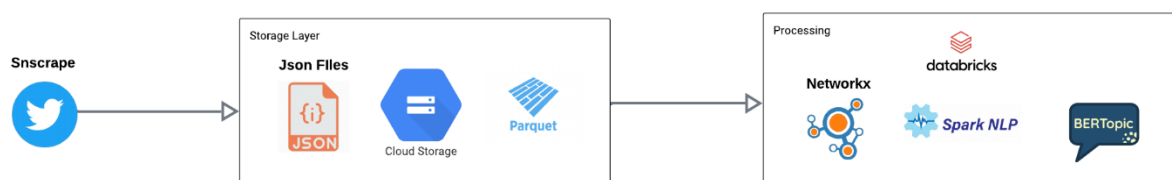


Figure 5 - Architecture Overview

4.2. DATA EXTRACTION

This project will focus on the two main presidential candidates in Brazil’s presidential election of 2022, Luis Inácio Lula da Silva, and Jair Messias Bolsonaro. From now on, for simplicity, they will be referred to only as Lula and Bolsonaro.

The first step was to extract tweets from the two candidates. Considering that the elections took place in October of 2022, the search criteria were composed of tweets published by Lula or Bolsonaro from January to October of 2022. Using Sns scrape’s CLI the tweets were extracted and saved in JSON file format in the GCS bucket. The full Sns scrape commands used in this project can be found in Appendix 9.1.

In the end, 3,741 tweets from Lula were extracted, and 1,871 from Bolsonaro. Figure 6 displays the number of tweets per month for each candidate. Lula produced more content than Bolsonaro, especially in the months close to the election.

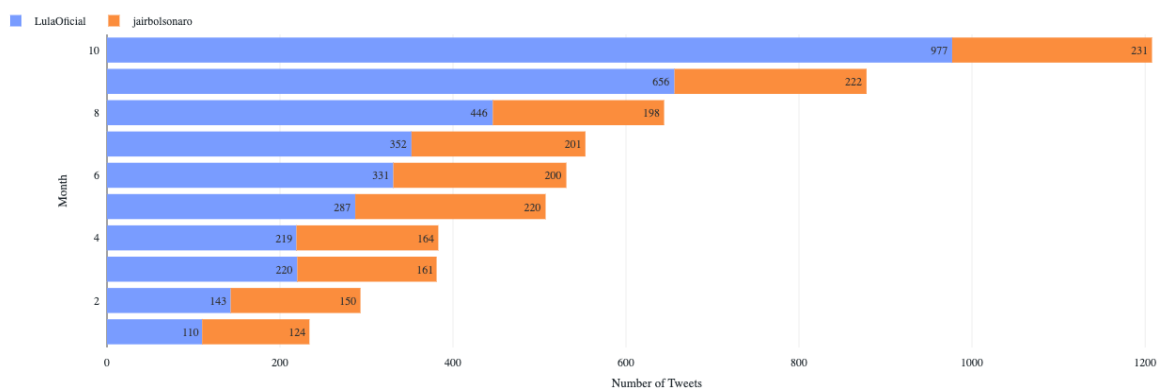


Figure 6 - Tweets from the Candidates

The second step was to extract all tweets which include mentions and replies to the given candidates. See Appendix 9.1 for a complete description of the sns scrape commands used. The search criteria were tweets that include mentions to one of the two candidates between January and October of 2022. 13,448,074 tweets from 1,497,514 users were retrieved with this query.

Data extracted from Twitter contains different information including data about the users mentioned in the tweet, profile information about the user who create that tweet, and basic stats like the number of replies, number of retweets, etc. Not all the fields available will be used in the scope of this project. The dataset schema with the selected fields is presented in Table 1. The most important field is the content, of course, containing the text in its raw form with special chars, emojis, URLs, and other non-textual information. Another important property is conversationId that can be used to indicate which tweet starts a particular thread. Is useful to identify replies to a particular tweet and even continuations of a particular subject. The user property is an object with the user’s Twitter profile including some stats like the number of followers, number of friends, etc.

Table 1 - Dataset Schema

Field	Description
id	Unique identifier of a Tweet
url	Url of the Tweet
date	Date where the tweet was posted
content	The raw text of the tweet
user	User Profile containing his username and location
replyCount	Number of replies in this tweet
retweetCount	Number of retweets for this tweet
conversationId	Id of the replied tweet (for replies only)
lang	The language of the tweet
hashtags	List of hahstags used in the tweet
mentionedUsers	List of users tagged in the tweet

Not all the users had the location information in their profile. Also, there is no predefined pattern, the users can type anything they want, so this information is not very complete. After doing some text preprocessing the location was extracted from 67,767 users. As is possible to visualize in Table 2 (São Paulo, Rio de Janeiro, and Minas Gerais), the southeast region holds a huge part of the engagement. Also, among the top 10 locations Portugal and the United States (USA) are highly active. This can be explained by the number of Brazilian ex-pats living in both countries.

Table 2 - Top 10 replies by Location

Location	Total
São Paulo	17,273
Rio de Janeiro	12,531
Minas Gerais	6,320
Pernambuco	4,969
Portugal	3,589
Rio Grande do Sul	3,476
Paraná	3,073
Bahia	2,976
Santa Catarina	2,724
United States	2,650

4.3. TOPIC MODELING

The purpose of this task is to cluster candidates into meaningful topics that can provide a clear understanding of each candidate platform. This is not a trivial task, considering that tweets have a 140-char limit, and excluding non-textual information, this number can become even low. Small text data can have a negative impact on the results.

Also, Topic Modeling is very subjective in terms of evaluation. Existent evaluation measures can help to reach a conclusion about the best model and decrease your search space in terms of choosing the number of clusters, but they do not guarantee that it would have the best results. Topic Modeling still depends on the human eye to evaluate the results.

In the following sections, a detailed description of the modeling phase will be discussed. Starting with the Data Preprocessing step where text from the candidate's tweets will be prepared to be used as the model input. Next, different Topic Modeling algorithms will be tested to choose the one that can produce the best results. Finally, the selected model will be trained to select the topics.

4.3.1. Preprocessing

Natural Language Processing (NLP) is an area of research that provides techniques to understand and manipulate natural language text using computer programs. NLP compasses a variety of disciplines and study fields, such as information sciences, linguistics, mathematics, psychology, machine translation, speech recognition, etc., to extract useful information from social media data, it is extremely important to clean up and prepare the data. Especially when working with social media where data may have a lot of non-textual information like images, videos, and emojis. Preprocessing is a very important NLP task used to clean up and prepare the text for further steps [27].

Just like in any other Machine Learning task, in a Topic Modeling approach, the data quality is the key to achieving satisfactory results. Using social media data and especially Twitter as input data makes this task even more difficult. Social media data contains slang, abbreviations, non-text data (emojis), videos, and pictures.

Common Python libraries and regular expressions can be used to preprocess text data. However, when data starts to grow traditional libraries are not so efficient as they will consume too much memory and will take a long time to run. For big datasets, a good approach is to distribute this process between several machines using a distributed computing technology like Spark NLP ¹. Spark NLP is an NLP library built on top of Apache Spark ML. Spark NLP provides several NLP and machine learning tasks that can scale easily in a distributed environment. It provides state-of-the-art models and libraries not only for Python but for other different programming languages [28]. Spark NLP can provide scalability from an NLP project by leveraging the Apache Spark engine. In this project Spark NLP will be used for the data preprocessing step.

In Spark NLP, data preprocessing tasks are called annotators. Annotators are grouped to form an NLP Pipeline, which is a sequence of transformations applied to a given corpus. The Pipeline can be serialized and later reused again by any project. In this project, a Spark NLP Pipeline was created to preprocess the data.

¹ <https://nlp.johnsnowlabs.com/>

Text data from tweets contains a lot of special characters especially considering the common signs to mention a user (“@”) and the hashtag sign (“#”). Also, any other characters, emojis, and non-textual data users may share are not useful for the Topic Modeling algorithm. Therefore, the first step was to clean up any occurrence of these characters. Spark NLP provides the DocumentNormalizer annotator that was used to clean up text based on predefined patterns. This annotator will automatically remove any special characters and leave only words in the text.

To preprocess the text is particularly important to break it into meaningful parts and identify the individual entities present in the sentence. This process is named Tokenization and is usually executed at the beginning of the preprocessing step. Tokenization is performed by identifying boundaries in the text such as spaces and punctuation and splitting the text based on these boundaries [29]. The Tokenizer annotator splits the words in a text and creates the tokens.

Every language can have common words that can frequently occur in a text (e.g., articles, prepositions, pronouns, conjunctions, etc.). In NLP, these common words are called stop words and they should be removed since they can increase the dimensionality of the term space. Apart from that, stop words usually do not add significant information to most of the NLP tasks and can create some noise [29], [30]. The common way of removing stop words is by using a predefined stop words list for the given language and removing them from the text. In Spark NLP, the StopWordsCleaner annotator is used to remove the stop words.

To extract meaningful topics, it is a good practice to focus on grammatical structures that can clearly describe each topic. Prepositions, adverbs, pronouns, and verbs clearly cannot provide a good understanding of the topic behind a particular text. Nouns and adjectives or a combination of both can provide a clear understanding of the subject behind the topic. To extract only these words a Part of speech (POS) tagging annotator was used. Part of speech (POS) tagging is the process of assigning a label (or tag) to each word in a sentence. In many NLP tasks such as information recovery, information handling machine interpretation, grammar checking, machine translation, and automatic summarization, POS tagging is reflected as one of the most necessary tools. Usually, a standard set of tags is selected to work with. Each tag corresponds to one grammatical structure like verbs, adjectives, nouns, etc [31]. The annotator was configured to select only nouns, adjectives, or a sequence of both.

Some words can have different meanings or add more contextual information if they appear together with other words. Those are called n-grams and can be described as a sequence of two or more adjacent words extracted from a section of continuous text. N-grams are frequently used to represent features in a vector space. A common convention is that “n” in n-grams corresponds to the number of elements in the sequence [32]. The NGramGenerator annotator was used to select unigrams and bigrams from the corpus.

Political tweets contain a lot of words that do not add significant meaning to the topics. Person names, cities, states, and users. In the preprocessing step, these words were removed so the algorithm can focus on other words. A list with common Brazilian names ² was used to remove person names from the text. In the same way, lists with all the cities and states ³ in Brazil were used to remove these words from the text.

It is also important to deal with Topic General Words (TGWs), which are words that, due to their higher frequency, can impact the output [33]. TGWs and other words that created noise in the topic

² From: <https://gist.github.com/augustohp/2c59ceb96e195ea375abadb311637e7f>

³ From: <https://raw.githubusercontent.com/datasets-br/state-codes/master/data/br-state-codes.csv>

modeling output were removed during the modeling stage by looking at the model output. Words that appear in several different topics as the most frequent ones were added to the stop words list, and the preprocessing and model training step was executed again until a good output was generated.

The final Spark NLP pipeline can be found in Appendix 9.2.

4.3.2. Model Selection

Latent Dirichlet Allocation (LDA) was proposed by Blei et al. [15] and is the most popular TM algorithm. LDA is a generative probabilistic model of a corpus. LDA works on a premise that documents are represented as a random mixture of topics, and each topic is a distribution of words.

LDA assumes that each document (M), which is composed of a set of words (N), can be represented as a probabilistic distribution of Dirichlet on latent topics [35]. In Figure 7 [36] is possible to observe the operation of the LDA model:

- α represents Dirichlet prior weight of the topic by document;
- Z represents the process of assigning a word to a topic,
- W represents the word in document M

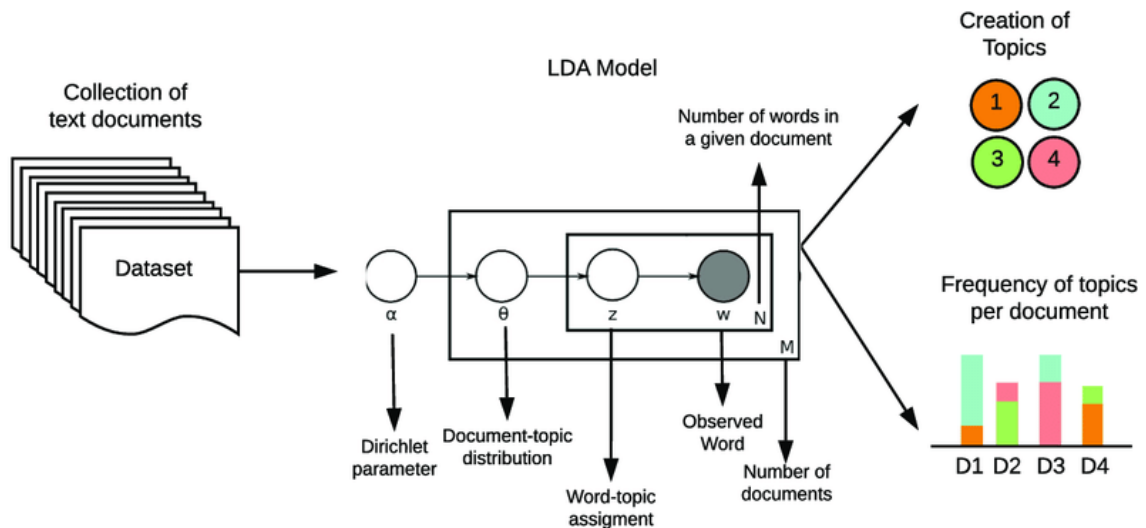


Figure 7 - LDA Model Operation (Buenano-Fernandez D. et al.)

In summation, for LDA a document is composed of a distribution of topics and a topic is composed of a distribution of words. In the end, LDA will provide the percentage of contribution of each topic to a particular document. This is a good approach, especially for big documents that can have several paragraphs discussing different subjects. For short texts, however, this is not true.

LDA was the first model tested in this project but did not produce reliable results. The model was implemented in Python and the Coherence Score ([37], [38]) was used to evaluate the results. Figure 8 displays the Coherence Score for each number of clusters for this experiment. The best value was for 9 clusters but is a huge value (0.98), which is a good indication that something is wrong. The next best value was for 4 clusters. This number is a bit low considering the corpus and the different subjects discussed by each candidate. However, the coherence score is not the only factor to be

considered to evaluate a Topic Modeling output. The selected topics should be examined to understand how coherent they are.

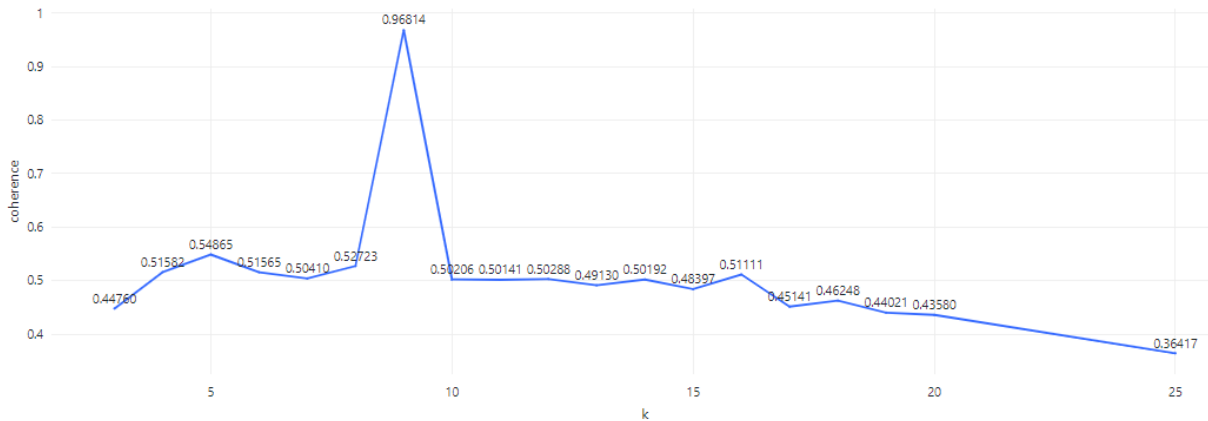


Figure 8 - Coherence Score x Number of Topics

Figure 9 presents the topics extracted by LDA. The generated topics don't reflect all the themes and discussions initiated by both candidates. Also, these topics are not clear enough to be interpreted and understand the subjects behind them. Apart from that, some words are present in different topics creating noise and impacting the analysis.



Figure 9 - Word Clouds for LDA Topic Model

An alternative to LDA that works better with short text data is GSDMM. Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture Model (GSDMM) introduced by Yin Ji and Wang Ji [39] is a topic modeling algorithm for cluster short-text data. The algorithm can be described using the Movie Group Process (MGP) analogy:

Imagine that a professor in a class of movie discussions wants to divide students into several groups. She would like students in the same group to have watched the same movies, so they can have more topics to discuss. Each student should then write a list of movies they watched, only the recent ones, so the list is not too long. The professor will then group the students based on similarities in their lists. Formally, GSDMM works as follows:

- The input is D students (or documents)
- Each student (or document) is represented by a list of movies (or words)
- Cluster students (or documents) in the same group by minimizing the differences between the list of movies (documents) from students in the same group and maximizing the difference between students from other groups.

To illustrate the assignment process of MGP the author proposed the following scenario:

“We can imagine that the professor invites the students into a huge restaurant and randomly assigns the students to K tables. Then she asks the students to re-choose a table in turn. We can expect that a student will choose a table according to the following two rules:

- *Rule 1: Choose a table with more students.*
- *Rule 2: Choose a table whose students share similar interests (i.e., watched more movies of the same) with him.”*

In this scenario, some tables will become full of students and others will be empty. Also, students that are already at a popular table will be likely to stay there. This analogy is equivalent to the Gibbs Sampling algorithm applied on GSDMM.

Like LDA, GSDMM also requires the definition of the number of topics to run. The Coherence Score was used again to identify the best number of topics. With a score of 0.5041, 5 topics were selected. Figure 10 displays the word clouds for the selected topics. GSDMM also did not produce superior results with this number of topics. Apart from having the best coherence score and the words seem not to repeat against topics this number of topics does not reflect reality.



Figure 10 - Word Clouds for GSDMM Topic Model

To improve the results with GSDMM the correct number of topics needs to be defined. However, testing several different values is not very practical since there is no way to select this number based on the metric. The output must be analyzed to evaluate the model.

To overcome this problem the solution was to use a model that does not require the number of topics: BERTopic [40]. BERTopic leverages clustering techniques and a class-based variation of TF-IDF to generate coherent topic representations. Using a pre-trained language model, document embeddings are generated to create a semantic representation of the document. BERTopic works in three steps described as follows:

1. Document Embeddings: The first step is to convert each document to its embedding representation using a pre-trained language model. The model assumes that documents from the same topic are semantically similar. Any embedding technique can use for this purpose if the language model was fine-tuned on semantic similarity.
2. Document Clustering: To reduce the dimensionality space BERTopic uses UMAP since it can be used across language models with different dimensional spaces. The reduced embedding is clustered using HDBSCAN, a version of DBSCAN changed to work as a hierarchical

clustering to create the clusters. HDBSCAN helps prevent unrelated documents being assigned to any cluster by modeling them as outliers and improving topic representation.

3. Topic Representation: The topic representation is defined based on a generalized version of the TF-IDF algorithm by concatenating all the documents within a cluster into one document only. With the TF-IDF matrix ready the model can generate a topic-word distribution to represent each topic based on the most important words.

BERTopic does not require the definition of the number of clusters before his execution. It will automatically choose the best number of topics and will also identify outliers. If required, it is possible to set the number of clusters by replacing HDBSCAN with K-means as the cluster model. The drawback of this approach is that K-means do not generate outliers, and this can cause some documents to be assigned to the wrong topics.

In this project, BERTopic was first used with HDBSCAN to select the topics. This experiment produced 125 topics with 1720 documents classified as outliers. Apart from having meaningful topics at this point, it was also clear that some topics could be merged as they were part of a more high-level subject. In Figure 11, is possible to observe that the first three topics can be merged into one topic since they are all about education. The other two topics are about fake news and lies and they also can be merged.



Figure 11 - Similar Topics that can be merged

This number of topics is also difficult to interpret and analyze. To overcome that, the documents in the outliers were removed from the initial corpora and the model was executed again using K-means as the cluster model and the number of topics.

The coherence score was used to select the optimal number of topics. According to RODER et al. [37], coherence is not an absolute measure of topic quality and values can vary depending on the corpus or domain. Nevertheless, in his work, he found that the best scores found by coherence measures lie in a range of 0.4 and 0.6. In this case, 17 topics were selected, with a coherence score of 0.47 producing satisfactory results (Figure 12).

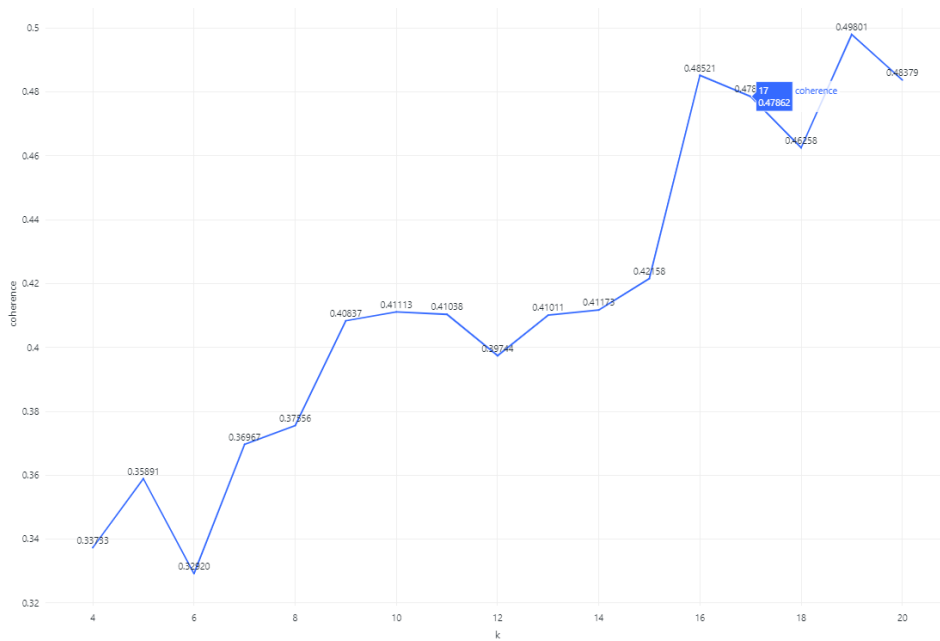


Figure 12 - Coherence Score for BERTopic

BERTopic output contains the list of topics, the list of documents assigned to each topic, and the set of probabilities for each assignment. Using this output, the replies for each tweet were assigned to the topic corresponding to the tweet they were replying to. Figure 13 displays the word clouds for the selected topics.



Figure 13 - Word Clouds for BERTopic

Appendix 9.3 contains some representative documents (tweets) per topic.

4.4. ENGAGEMENT GRAPH

Putnam presented the concept of social capital, which is the idea of the value provided by a social network. The assumption is that civically engaged communities are more likely to produce positive outcomes. That assumption was supported by researchers from different fields such as education, urban poverty, and unemployment. Later, he introduced two types of social capital: bonding, which brings a group together, and bridging which divides a group. Based on this concept RAYMOND C. [6], proposed a theoretical framework to study social engagement based on bridging and bonding messages. The idea is that some users may interact with certain types of content when they are produced by a given producer and when this content changes, they choose not to interact.

To identify these messages an engagement graph needs to be created. An engagement graph is a special type of graph that have 3 different vertices:

- Producers: users who tweet about a certain topic
- Content: The tweet about this specific topic
- Consumers: General users who interact with this content

To build the engagement graph is necessary to prepare the vertices based on:

1. Unique candidates' usernames
2. Unique ids of tweets posted by candidates
3. Unique ids of the replies from a user to a candidate

The engagement graph can provide useful information about how users interact with certain topics given by the producer. Every time a candidate posts a tweet a new connection is created between the Producer and the Tweet. When a general user replies to a tweet a connection is created between this user and the tweet. All the graph representations and algorithms used in this project were implemented using the Networkx library [41]. The commands used to create the network can be found in Appendix 9.3.

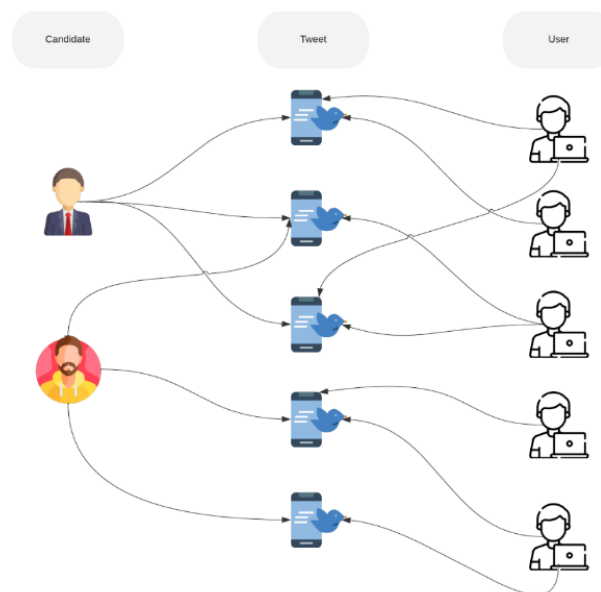


Figure 14 - Engagement Graph Representation

Given the number of nodes in the network, a full engagement graph cannot be visualized. Figure 15 illustrates the Engagement Graph for a subgraph of users with the location identified. It is possible to see 2 big nodes in the center of the graph representing each candidate: the green one is Bolsonaro and the pink one is Lula. Also, it is clear that some users only interact with one of the candidates, and a small portion interact with both.

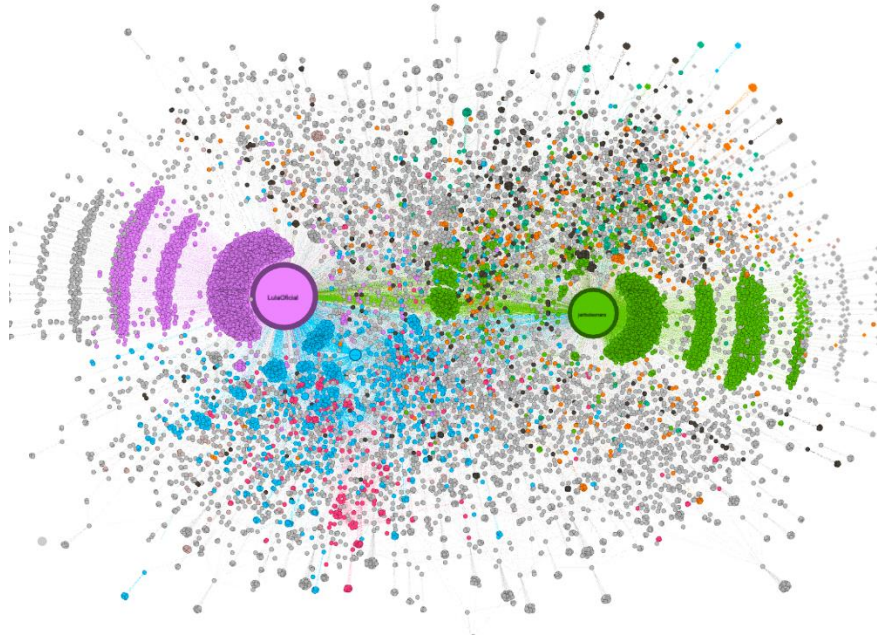


Figure 15 - Engagement Graph

After creating the Engagement Graph the next step is to compute the measures of Topic Centrality. There are many ways to measure how central a node is in a network. The easier way is just by counting the number of connections the node has. However, some methods can be more efficient since they not only measure the number of connections but also how active the connected nodes are in the network.

Eigenvector Centrality is a measure of graph centrality where the centrality of a node is proportional to the average centrality of its neighbors [34]. The Eigenvector Centrality will be used to compute the measures of Topic Centrality with the following steps:

1. Start by computing the Eigenvector Centrality for each tweet in the network.
2. Compute the Total Network Topic Centrality by averaging the Eigenvector Centrality for each tweet and reply in the network.
3. Candidate's Network Topic Centrality: Computed for each candidate. Start by removing all the tweets and replies from the other candidate from the network and then computing the Eigenvector Centrality for the remaining tweets and aggregating them.

After that, these measures will be compared to understand which topics are driving more engagement for each candidate's network. High Topic Centrality values will indicate that this topic was replied to by highly active users in the network. These users are very engaged in the network and therefore participate in different discussions. If this behavior is observed in the Total Network Topic Centrality this also indicates this topic has overall importance and can reach users from both communities.

4.5. SENTIMENT ANALYSIS

Polarization refers to a broad concept that can compass diverse types, such as ideological and affective. Frequently, in the literature, polarization is studied from an ideological point of view, where users` disagreement happens because of their different opinions around a particular subject. Nevertheless, affective polarization is also commonly observed in real life. People tend to vote for or support candidates they like, regardless of their actions or opinions.

In this step, the sentiments of users around topics and candidates will be measured. This task will be performed with the following steps:


1. Extract sentiments from tweets: In this step, the sentiment from each tweet will be extracted and the polarity score will be computed.
2. Aggregate sentiments per Topic: The scores selected in the previous step will be aggregated by user and topic to select the User Polarity Score per Topic.
3. Select Hashtags: Hashtags containing the name of each candidate will be extracted.
4. Extract sentiments from hashtags: Compute the polarity score for each hashtag.
5. Aggregate sentiments from Hashtags: Hashtags will be grouped by candidate and polarity score.

The LeIA python library was used to extract the sentiments for each tweet, including the replies, in the dataset. LeIA (Léxico para Inferência Adaptada) [43] is a fork from the VADER (Valence Aware Dictionary and Sentiment Reasoner) [38] lexicon adapted to Portuguese. LeIA supports emojis and is focused on social media texts. LeIA produced an output with the following fields:

- pos: percentage of positive sentiments in the text
- neg: percentage of negative sentiments in the text
- neu: percentage of neutral sentiments in the text
- compound: normalized value from -1 to +1.

In this case, the compound will be used to identify the sentiment polarity of each tweet. Table 3 presents some examples of polarities identified by LeIA.

Table 3 - Sentiments Polarity on Tweets

text	neg	pos	neu	compound
@brom_elisa @jairbolsonaro Misericórdia, isso é mal de família ou é a família do mal?	0.492	0.061	0.447	-0.8658
@simonetebetbr @LulaOficial Parabens senadora!! Sua ajuda contou demais!!! 	0.0	0.561	0.439	0.8527
@FlavioBolsonaro @jairbolsonaro Flavinho, quantas vezes eu já te disse? MOTO NÃO VOTA!!!	0.257	0.0	0.743	-0.5871

After classifying the sentiment for all tweets, the sentiments expressed in hashtags were computed. The criteria for selecting the hashtags were entries that contain the name of any candidate, or any variation of it. For example:

- For Lula: Any hashtag containing the words “Lula”, “LuisInacioLulaDaSilva”
- For Bolsonaro: Any hashtag containing the words “Bolsonaro,” “JairBolsonaro”, “JairMessiasBolonaro”, “Bolso”

Using these criteria, 4,085 hashtags from Lula were extracted, and 2,914 for Bolsonaro. Apart from that, 112 hashtags containing names from both candidates were extracted. The words in the hashtags were extracted using regular expressions to use them as input to the LeIA model. For this task, the tweet's raw content was used since LeIA supports emojis and works well with social media text. Figure 16 presents the most used hashtags.

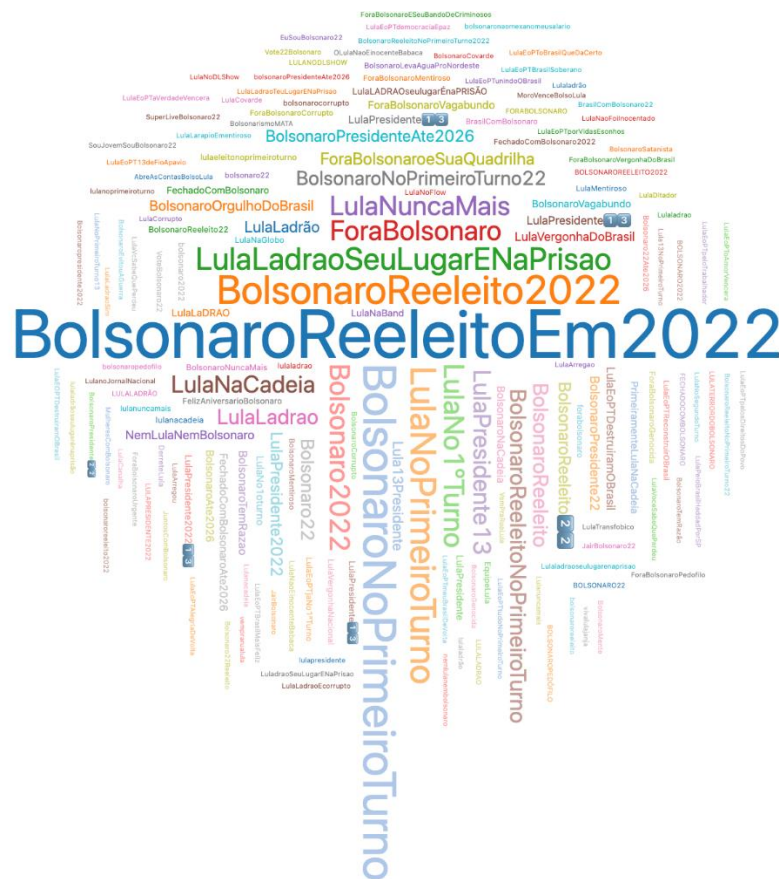


Figure 16 - Word Clouds Hashtags

Finally, 5 groups of hashtags were defined according to the polarity scores in the hashtags:

1. Pro-Lula: Hashtags mentioning Lula with a positive score
2. Against-Lula: Hashtags mentioning Lula with a negative score
3. Pro-Bolsonaro: Hashtags mentioning Bolsonaro with a positive score
4. Against-Bolsonaro: Hashtags mentioning Bolsonaro with a negative score
5. Neutral: Hashtags with terms the indicates users do not support any candidate

5. RESULTS AND DISCUSSION

So far, some interesting features were extracted from the dataset:

1. 17 Topics from the Topic Modeling Algorithm
2. An Engagement Graph from a network of tweets and replies
3. The Measures of Topic Centrality from the Engagement Graph
4. Sentiments from tweets and hashtags

In this section, the extracted features will be combined to understand how users interact with candidates and which factors can contribute to polarization. By looking at the Topic Centrality measures the goal is to understand if users' engagement increases or decreases concerning the topic. The sentiments extracted from the tweets will help to understand if the shared content drives positive or negative sentiments from users. Finally, hashtags will give a more direct definition regarding users' feelings toward candidates.

First, users will be segmented into 3 distinct communities according to their interactions with the candidates:

1. Users who only replied to Bolsonaro's tweets
2. Users who only replied to Lula's tweets
3. Users who replied to tweets from both candidates

These communities will be used to expand the analysis of engagement behavior and sentiments toward candidates.

Figure 17 presents the distribution of users in each community. Almost 50% (48.8) of the users interacted only with Lula, 30.1% exclusively with Bolsonaro, and 21.1% of the users interacted with both candidates.

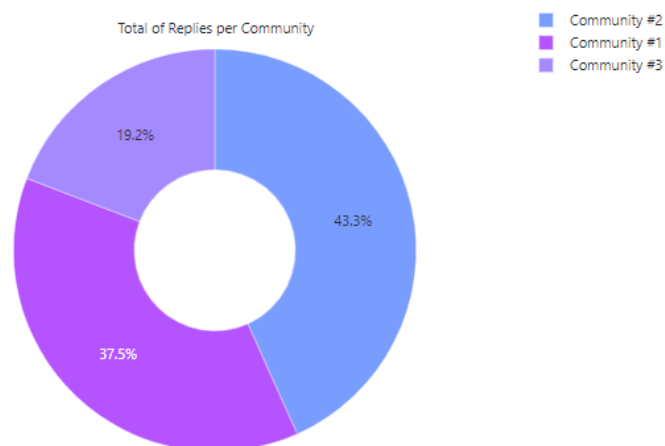


Figure 17 - Distribution of users per Community

5.1. TOPIC MODELING

After creating the model, 17 topics were selected from the candidate's tweets. Each topic has a set of keywords that are more meaningful for that topic and can be used to interpret it. The selected topics are presented below together with the keywords for each one:

- Topics 0 (fé, obrigado, carinho): This topic contains messages where the candidates express their gratitude towards the support they receive from people and their faith in helping the country.
- Topic 1 (caminhada, grande, ruas): This topic is about campaign rallies and other events.
- Topic 2 (Fake News, mentiras, verdade): It has several messages about fake news and contains mutual accusations of spreading fake news.
- Topic 3 (Fome, comer, comida, alimentos): About fighting hunger and decreasing the taxes on food and agriculture.
- Topic 4 (democracia, mundo, paz): Democracy is the subject discussed in this topic.
- Topic 5 (radio, conversa, imprensa): Topic about candidate's interviews on radio, tv, and other platforms.
- Topic 6: (prefeitos, governadores, eleições): Given that the year of the presidential election is also the year of the governor's election, topic 6 candidates express their support to governor candidates.
- Topic 7 (salário, inflação, orçamento): This topic contains discussions about the country's economy and themes like: social inequality, minimum wage, and inflation rate.
- Topic 8 (educação, universidades, escolas): Several discussions about education including how the candidates pretend to improve the education system in Brazil.
- Topic 9 (água, transposição): The discussions in this topic are about sewage disposal, and bringing water to cities that suffer from the drought, like the transposition of the São Francisco River, which was a big project of conducting the waters to the northeast watersheds.
- Topic 10 (cultura, futuro, juventude): Candidates talk about making the country a better place for children and young generations by making education, culture, and basic needs available for them.
- Topic 11 (casa, família, vida): This topic is about programs to build houses so everyone can have a place to live.
- Topic 12 (gasoline, preço, redução, diesel): This topic holds discussions about the reduction of fuel prices.
- Topic 13 (mulheres, brasileiras, homens): In this topic opinions about violence and discrimination against women and other minorities are discussed.
- Topic 14 (seguro, apoio, números): In this topic, candidates express their gratitude towards the support of people in campaign events.
- Topic 15 (quinta-feira, registrar, presença): Another topic about campaign events. Candidates are inviting people to the events, especially the live streams, and for them to sign up on their social media platforms.
- Topic 16 (amazônia, indígenas, povos): This topic is about the Amazon Forest and native people and the recent discussions about deforestation, fire, and other issues that impact their life.

Looking at the distribution of topics per candidate (Figure 18), Lula was more active than Bolsonaro on Twitter. Excluding topics related to campaign activities (0,1,5,14) the Lula produced more content related to topics 2 (Fake news and lies), 3 (Hunger, food taxes, agriculture), and 4 (Democracy). Bolsonaro focused his content on topics 12 (Reducing prices and taxes on fuel), 9 (Sewage disposal infrastructure), and 7 (Social Inequality, minimum wage).

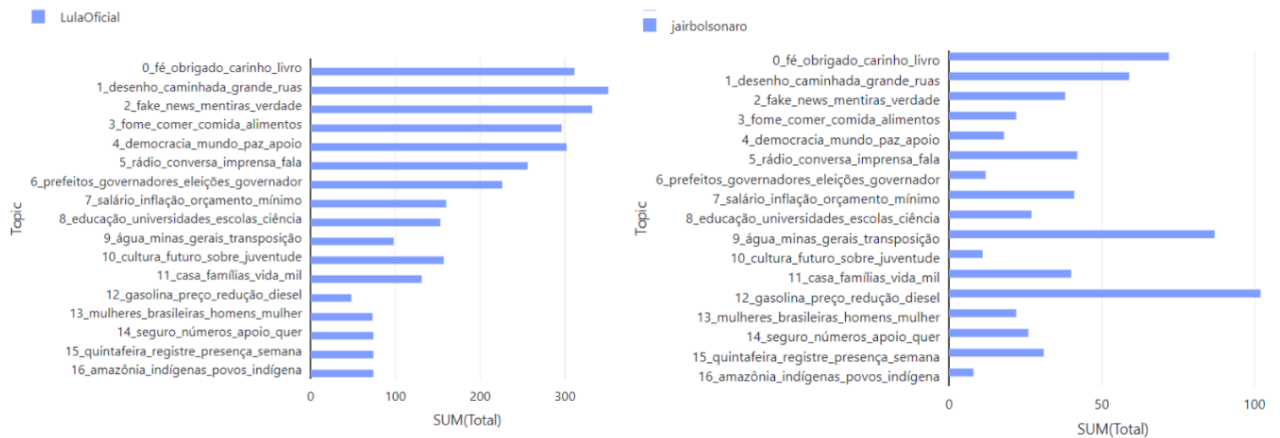


Figure 18 - Distribution of Topics Per Candidate

Figure 19 displays the number of replies per topic. Topic 2 is the one with more replies followed by 3 and 4. These topics represent some important discussions that took place in Brazil during the election:

- Topic 2: Candidates' mutual accusations of spreading fake news were constant during the last years. Bolsonaro also frequently accused the media of propagating fake news about his government.
- Topic 3: One of the pillars of Lula's campaign was his programs to end hunger in poor regions of Brazil. He argues that in his previous mandate, he was responsible for bringing food to poor families and that he will do that again once he became president. On the other hand, Bolsonaro's supporters argue that this was not the case, and even after his government, many people were still suffering from hunger.
- Topic 4: Democracy was also a frequent theme in these discussions. Both sides accused the other of creating an anti-democratic environment and that liberty of expression will be threatened by the opposing side.

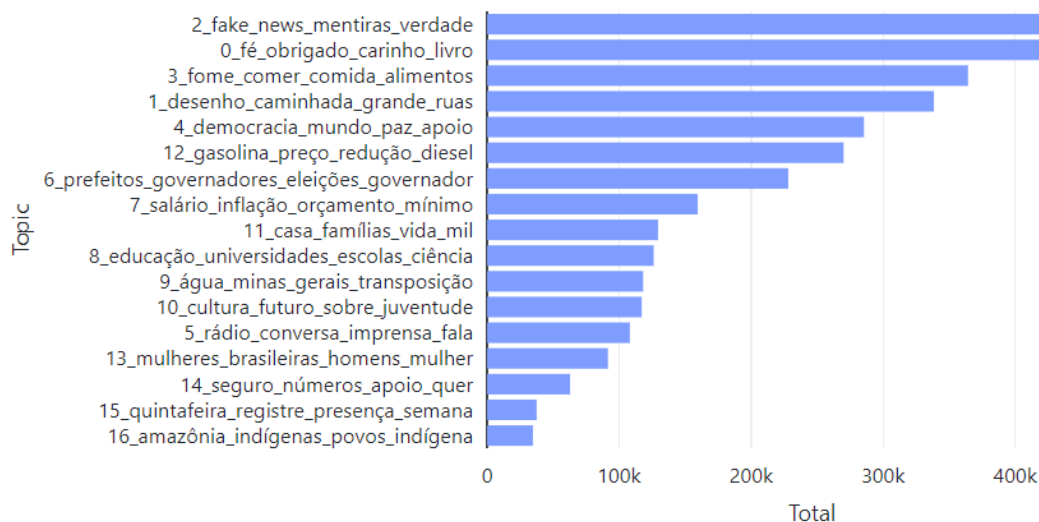


Figure 19 - Number of Replies Per Topic

5.2. TOPIC CENTRALITY

The Topic Centrality measures extracted from the Engagement Graph can help understand which topics can drive engagement for each candidate and the entire network. Figure 20 displays the Topic Centrality per topic for the entire network. Topics related to campaign activities were removed to simplify the analysis. It is possible to observe that Topics 2, 3, 4, and 12 are the more central topics in the network. These topics are the most frequent topics for both candidates and therefore also the ones with the most replies. This indicates that messages posted about these topics reach the entire network (or a big part of it). On the other hand, topics 5, 13, and 16 have the lowest Topic Centrality score and therefore they drive less engagement in the network. Topics 2 and 3 also have good importance for Lula's network but are not so important to Bolsonaro's network. Nevertheless, topic 12 is central to Bolsonaro's network opposing Lula's.

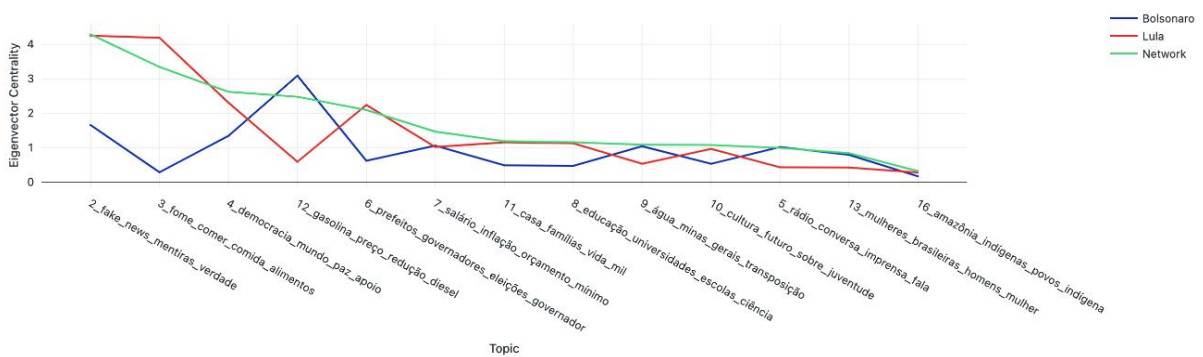


Figure 20 - Topic Centrality per Community

A scatterplot is a useful tool to compare the Topic Centrality Measures by plotting the values of two different samples on each axis. The upper left corner corresponds to topics that are more to the sample on axis y and less central to the sample on axis x. The lower right corner corresponds to the opposite: topics that are more central to the sample on axis x and less central to the sample on axis y. The upper right corner corresponds to topics that have a higher for both samples.

Figure 21 presents a scatter plot for the Network's Topic Centrality on axis y and Bolsonaro's Topic Centrality on axis x. As is possible to observe there are no huge discrepancies. Topic 3 seems to be a little more important to the network than the candidate and Topic 12 is a bit more important to the candidates. These two topics are indeed the most important ones for Lula's network (Topic 3) and the most important one for Bolsonaro's network (Topic 12). To reach a more diverse audience Bolsonaro could focus on topics 2, 3, and 4.

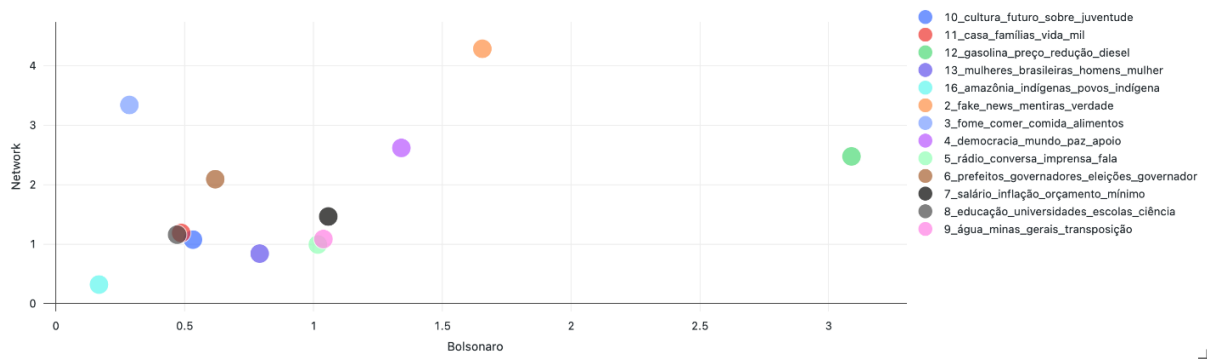


Figure 21 - Total Network's Topic Centrality x Bolsonaro's Topic Centrality

Figure 22 displays a scatterplot of Lula's Topic Centrality and the Network's Topic Centrality. In this case, the Topic Centrality for the candidate is close to the network values. Topics 2 and 3 are close to the upper right corner and therefore are important for the entire network and the candidate's network. A slight difference can be observed on Topic 12, which is a bit more important for the entire network than for Lula's. In this case, focusing on topic 12 will drive a more diverse audience.

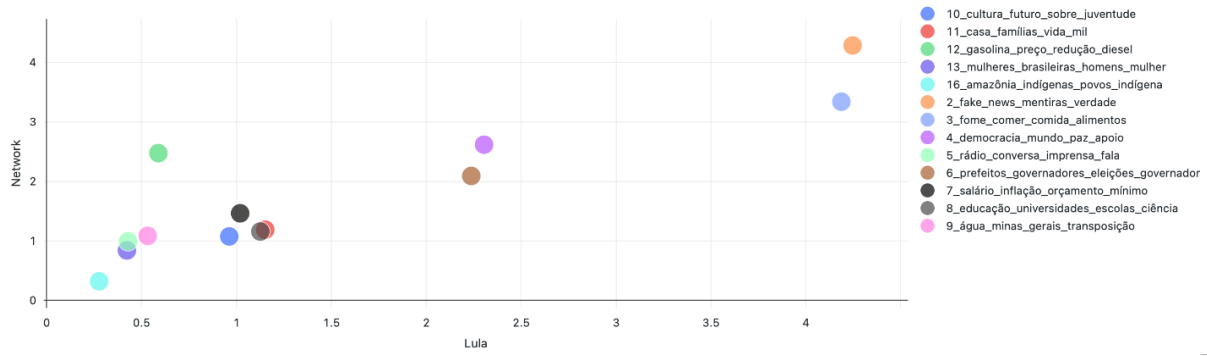


Figure 22 - Total Network's Topic Centrality x Lula's Topic Centrality

It is also interesting to compare the measures of both candidates. In Figure 23 it is possible to observe that in the upper left corner, Topic 12 is more central to Bolsonaro's network and less central to Lula's. This means that users do not interact so much with Lula when he retweets about these topics. In the same way in the lower right corner topics, 2 and 3 are more important for Lula's network than for Bolsonaro's.

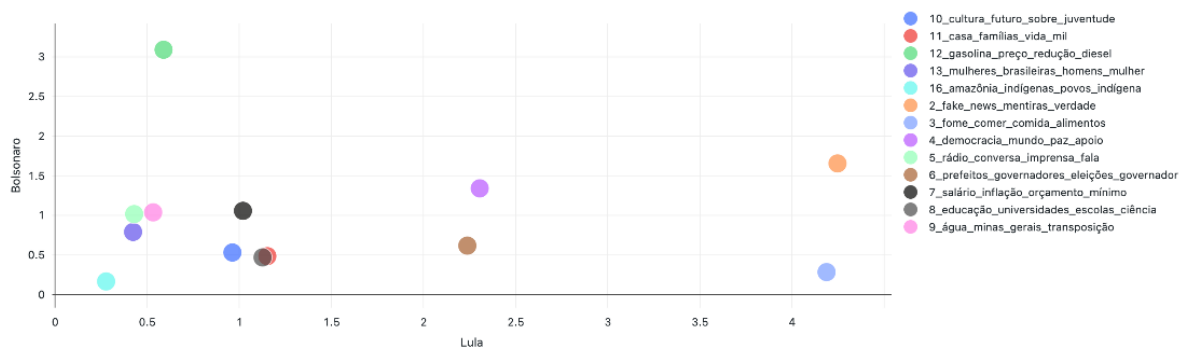


Figure 23 - Lula's Topic Centrality x Bolsonaro's Topic Centrality

These results show that only 3 topics can bridge or bond the communities. In general Topic 2 have a good reach as it is central to both, the Total Network and Lula’s Network. For Bolsonaro, it lies in the middle of the chart and therefore, this topic can potentially drive more engagement from both sides. Topic 12 for Bolsonaro and Topic 3 for Lula are the ones who can bridge the communities since they are the most central topics for each side and are not for the other side.

This behavior can be verified by looking at the replies per topic for the communities created at the beginning of this section (Figure 24). Topic 2 is the most replied topic for Groups 2 and 3 and is the second most replied topic for Group 1. This topic indeed has a good reach and can drive engagement across the groups as it is the most central topic for the entire network. Topic 3, on the other hand, has good traction in Group 2, from users who interact only with Lula, and in Group 3, from users who interact with both candidates. In Group 1, from users who only interact with Bolsonaro, the number of replies is low. The same thing happens with Topic 12 which has more replies from users in Group 1 than from the other groups. These two topics are bridge topics as they drive more attention from one community within the network.

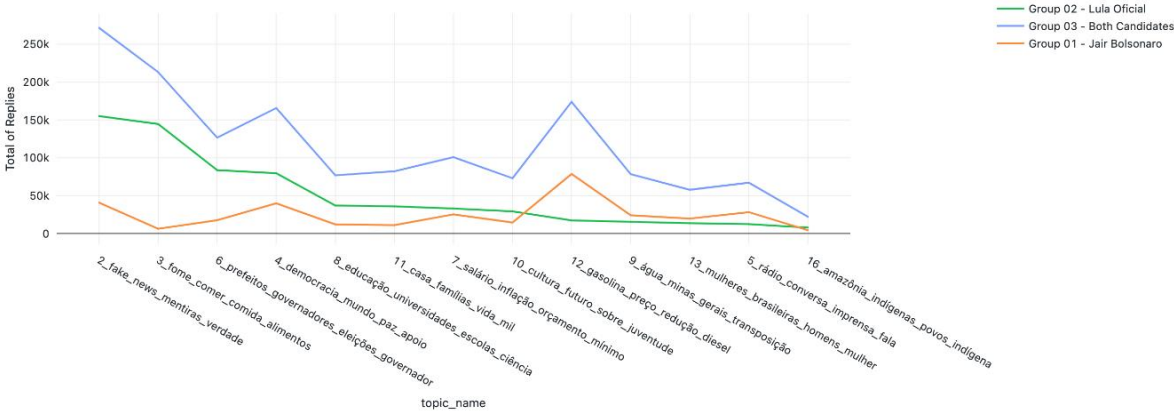


Figure 24 - Number of Replies per Topic and Community

These results are aligned with each candidate's ideological affinity. Bolsonaro is a right-aligned candidate, therefore topics related to the economy, like topics 12 and 7 have significant importance in this platform. Topic 12 is the most central topic for Bolsonaro and Topic 7 is the fourth one. Lula is a left-aligned candidate, therefore topics related to themes that decrease social inequality are more relevant to his platform. Topic 3 discusses hunger and is the second more central topic on Lula’s network. This is one of the strongest points on his platform since in his previous mandate the Zero Hunger Strategy was launched and was the subject of several discussions. Some author argues that this program has a unique potential for global diffusion [45].

5.3. SENTIMENT ANALYSIS

The predicted sentiments for each tweet were aggregated per candidate and general users. Figure 25 displays the average sentiment polarity for each topic. Looking at the candidate's scores (red and blue lines) it is possible to observe that the sentiment polarity varies according to the topic. For Bolsonaro, the most positive topic is Topic 10 which is about the future of children and the young population. The most negative is Topic 12 where he discusses the problems related to the prices of fuel. For Lula, Topic 11 is the most positive one and it is about his plans to provide houses to all the population. Topic 2 about fake news is the most negative topic from Lula.

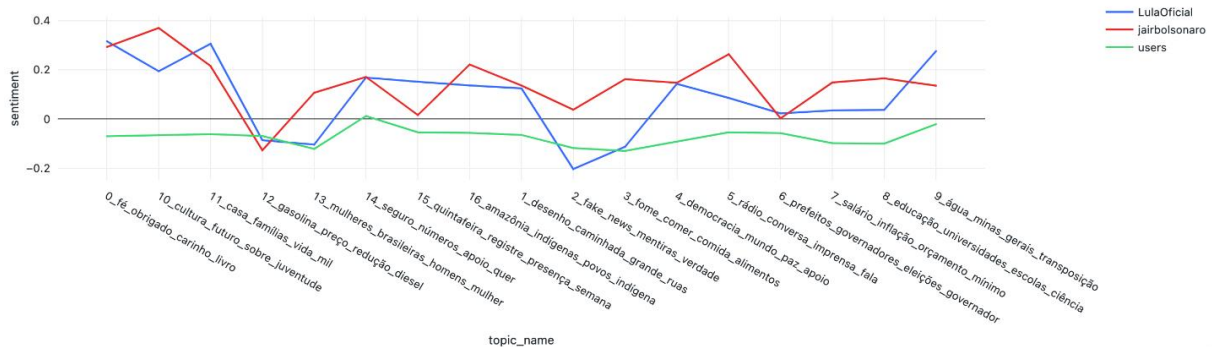


Figure 25 - Sentiments per Topic and Community

The user's sentiments do not show any significant variation according to the topic. As is possible to observe, the sentiments are always negative and constant across topics. This can be explained by the fact that users are more interested in sharing their disapproval of the candidate and their disbelief in his plans than expressing their support or providing good arguments that could enrich the discussions. In the users' replies to candidates' tweets, there were about 1,253,642 entries containing terms like "thief", "liar" and "fascist".

Comparing the sentiments within the communities with the candidates' sentiments the same pattern was found in extremely negative sentiments across the topics. Figure 26 illustrates this comparison, where the group's sentiments fall below the positive values.

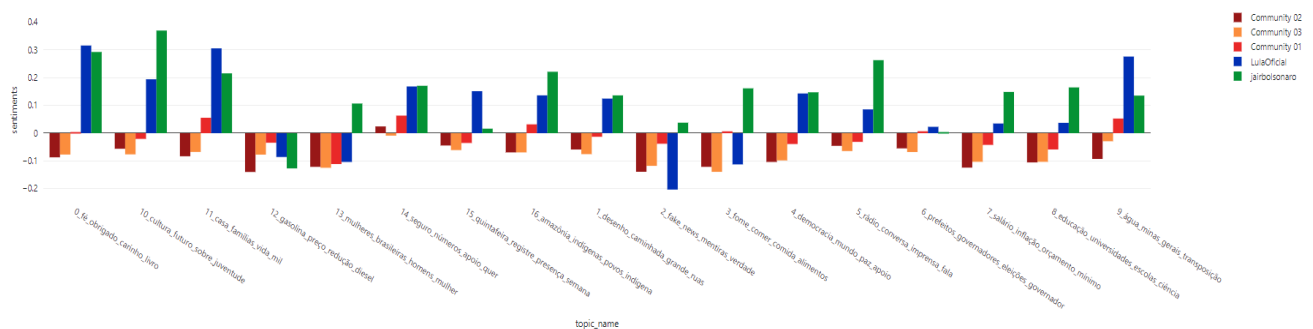


Figure 26 - Sentiments per Communities and Candidates

In summation, no variance across significant differences was found in user sentiments per topic for the entire network and even considering the defined groups. Users' sentiments tended to be negative, regardless of the topic or the candidate they interact with.

The computed sentiments from hashtags will be used to expand this analysis and evaluate the user's sentiments towards candidates. Table 4 displays the number of hashtags and the aggregated polarity for each one of the Hashtags Groups. This table also presents the top 3 most used hashtags per group. The positive hashtags (Pro-Lula, Pro-Bolsonaro) are a good measure of support and partisanship since in both cases users are expressing their confidence and the desire for their candidates' victory. The negative hashtags (Against-Lula, Against-Bolsonaro) contain intense and negative sentiments: in both cases, users believe the candidates are thieves that should be in jail. A small portion of the shared hashtags is from users that did not want any one of the candidates to win (Against-Both).

Table 4 - Distribution of Hashtags Per Group

Group	Top 3 Hashtags	# Total	Polarity
Pro-Lula	LulaNoPrimeiroTurno LulaNo1ºTurno LulaPresidente13	1829	0.93
Against-lula	LulaLadraoSeuLugarENaPrisao LulaNuncaMais LulaNaCadeia	2256	-0.82
Pro-Bolsonaro	BolsonaroReeleitoEm2022 BolsonaroNoPrimeiroTurno BolsonaroReeleito2022	1666	0.94
Against-Bolsonaro	ForaBolsonaro ForaBolsonaroeSuaQuadrilha BolsonaroNaCadeia	1248	-0.86
Against-Both	NemLulaNemBolsonaro AbreAsContasBolsoLula MoroVenceBolsoLula	112	-0.95

Another interesting measure is the percentage of partisan hashtags per candidate. Figure 27 displays these values, and it is possible to observe that in both cases users not only express their support for their candidate with positive hashtags by they are also active in expressing their disagreement with the other candidate. 28.32% of Lula's tweets have replies containing Pro-Lula hashtags and 30.53% have Against-Lula hashtags. A significant portion (36.5%) of his tweets contain Pro-Bolsonaro hashtags. For Bolsonaro, 53.07% of his tweets have replies with Pro-Bolsonaro hashtags and only 16.93% contain Pro-Lula hashtags.

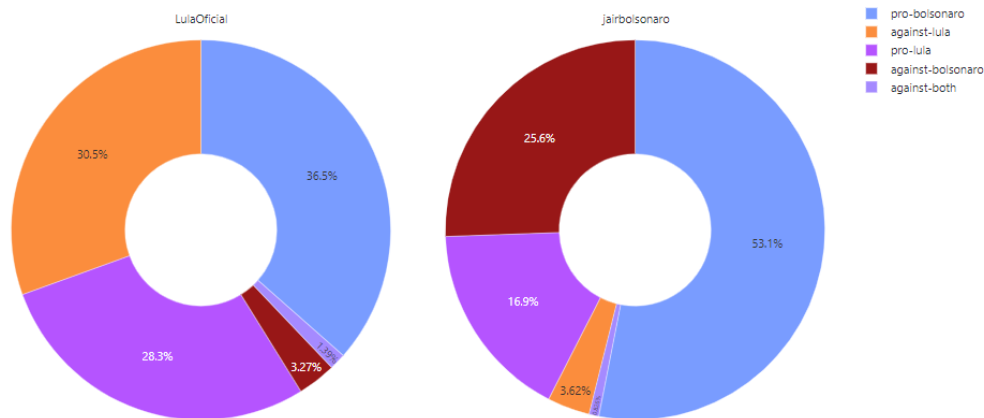


Figure 27 - Distribution of Hashtags per Group

The next step is to analyze the use of hashtags by communities to analyze how each one behaves in terms of expressing their support or their disagreement with the candidates. Table 5 presents the Top 5 most used hashtags for each community. They all have a combination of supporting and opposing hashtags which are curious since communities 1 and 2 have users who interact only with one of the candidates. For example, in Community 1, composed of users who only interacted with Bolsonaro, the Top 1 hashtag is ForaBolsonaro which means: “Get out Bolsonaro”. This indicates that some users that don’t support Bolsonaro were actively interacting with his posts just to express their disapproval. The same users however did not engage with Lula to express their support. Also, in Community 2 it is possible to observe that the hashtag BolsonaroReeleitoEm2022 is in the top 4 in this group. This again denotes that some users were more interested in engaging with Lula to express their support for Bolsonaro. In this case, some users were more interested in expressing their disregard for an opposing candidate than supporting their candidate.

Table 5 - Top 5 Hashtags per Community

Community	Top 5 Hashtags
Community 1 (Bolsonaro)	ForaBolsonaro ForaBolsonaroeSuaQuadrilha BolsonaroReeleitoEm2022 BolsonaroNoPrimeiroTurno Bolsonaro2022
Community 2 (Lula)	LulaNoPrimeiroTurno LulaPresidente13 LulaNo1ºTurno BolsonaroReeleitoEm2022 LulaLadraoSeuLugarENaPrisao
Community 3 (Both)	Lula2022 ForaBolsonaro BolsonaroReeleitoEm2022 BolsonaroReeleito2022 BolsonaroNoPrimeiroTurno

The percentage of hashtags per user was computed. Users were assigned to groups according to the percentages of hashtags they used. By doing that it was possible to visualize the sentiments each

group expressed about a topic. Figure 28 presents these values, and it is possible to observe that the most negative sentiments are in the against-both group. The pro-lula and pro-bolsonaro groups present the most positive sentiments.

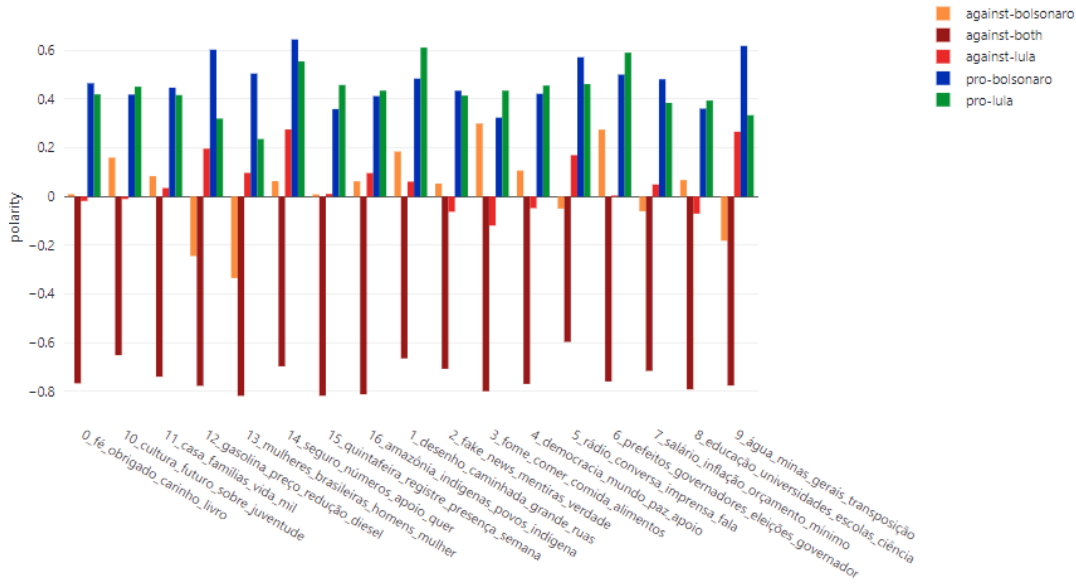


Figure 28 - Sentiments per Hashtag Group

Going further, it is possible to use the sentiments to measure the environment's polarization level. Users were again classified on two different spectrums (pro, against) according to the previous hashtag's groups. Users that are pro-lula and against-bolsonaro were classified as pro-lula only. Users that are pro-bolsonaro and against-lula were classified as pro-bolsonaro. A separate group was dedicated to against-both. Figure 29 presents the percentage of users on each spectrum. The majority of users lie on pro-bolsonaro or pro-lula and they represent almost 97% of the users in the population. Also, the difference between them is a little bit more than 7% with indicates a high level of polarization.

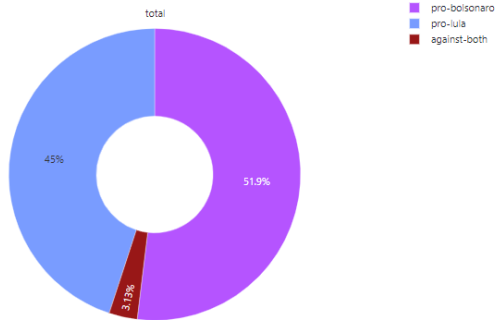


Figure 29 - Polarization Level

In summation, the results show that hashtags are a good asset to measure polarization and identify users` sentiments towards a particular entity and especially their partisanship in a political context.

6. CONCLUSION

This project analyzed the Political Polarization in Brazil based on data collected from Twitter regarding the presidential race in 2022. Tweets, mentions, and replies from the two main candidates, Jair Messias Bolsonaro, and Luis Inacio “Lula” da Silva, were collected and preprocessed. A Topic Modeling algorithm was applied to cluster the tweets in topics and 17 topics were selected and analyzed based on their importance. To find the most important topics for each candidate an Engagement Graph was built, and the measures of Topic Centrality were applied to the networks of both candidates. Finally, a Sentiment Analysis algorithm was used to identify the sentiments tweets text and hashtags.

The measures of Topic Centrality demonstrated that some topics have bonding and bridging capabilities:

- Topic 3: About fighting hunger, one of the pillars of Lula’s campaign is also the most central in his network. The centrality score for this topic on Bolsonaro’s network is low, so it has a bridging capability and drives more engagement for Lula.
- Topic 12: About fuel prices, is the more important in Bolsonaro’s network opposing to Lula’s. This topic drives more engagement to Bolsonaro.
- Topic 2: This is the most central topic for the entire network and therefore drives engagement from both sides. This topic has a bonding capability.

By analyzing the sentiments extracted from tweets it was clear that users in general express negative feelings when interacting with both candidates. In this case, no evidence was found that the content of the messages (topics) impacted the user’s sentiments. The candidates on the other hand have some positive sentiments on most of the topics. Lula and Bolsonaro presented negative feelings on both topics: Topic 2 (Fake news and lies) and Topic 12 (Hunger).

Sentiments extracted from the hashtags have proven to be extremely useful in identifying user partisanship. Users frequently use hashtags with terms that express their support or their disapproval of a candidate. This analysis also shows that some users choose to interact only with the opposing side to express their disapproval and do not engage with the candidate they support. Around 36% of replies to Lula’s tweets and 16% of Bolsonaro’s have hashtags containing messages from supporters of the opposing side.

Moreover, the results demonstrated that hashtags can be useful in measuring the level of polarization in an environment. It was found that around 51% of users are Lula’s supporters, 49% of the users are Bolsonaro’s supporters and 3.13% are against both candidates.

In summation, this study shows that users’ sentiments in candidate’s messages are mostly negatives and they do not vary according to the topic. The distribution of user sentiments per topic remained constant and negative for all the topics. Also, sentiments expressed in the hashtags indicate that some level of affective polarization since that they frequently use the hashtags in the posts to express their feeling about the candidates.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Extracting data from Twitter is difficult and sometimes not very productive. Twitter's API does not provide free access to the whole historical database. Snsrape helps in this case with CLI to extract historical data. However, since it is an open-source tool bug fixes are not fixed in a desirable timeframe. During the development of this project, the functionality of extracting profiles from Twitter was not working on Snsrape. This issue restricted this project since it was not able to extract the followers list from any user.

Retweets are also a good form of analyzing users' behavior on Twitter and can be used as a form of endorsement of the content. However, they are not easy to extract even using Snsrape, especially for historical data. In this case, they were not used in this project and only mentions and replies were considered.

Apart from that, both presidential candidates do not produce much content on Twitter. This can limit the use of some Topic Modeling algorithms as they depend on a good number of instances to perform well.

For future work, the recommendation is to compare topics and sentiments between the two last elections. In 2018, Bolsonaro, a right-wing candidate was the winner, and in 2022 a left-wing candidate, Lula was the winner. What can be the causes of this change in the voters' preferences? Is it possible to track this intention earlier?

Also, generalize this approach to detect, analyze, and understand polarization in other domains different from politics.

8. REFERENCES

- [1] C. Machado *et al.*, “News and Political Information Consumption in Brazil: Mapping the First Round of the 2018 Brazilian Presidential Election on Twitter COMPROP DATA MEMO,” 2018.
- [2] C. S. Park, “Does Twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement,” *Comput Human Behav*, vol. 29, no. 4, pp. 1641–1648, Jul. 2013, doi: 10.1016/J.CHB.2013.01.044.
- [3] Brent Kitchens, Steven L. Johnson, and Peter Gray, “UNDERSTANDING ECHO CHAMBERS AND FILTER BUBBLES: THE IMPACT OF SOCIAL MEDIA ON DIVERSIFICATION AND PARTISAN SHIFTS IN NEWS CONSUMPTION,” *MIS Q*, vol. 44, no. 4, pp. 1987–2011, Aug. 2020, doi: 10.25300/MISQ/2020/16371.
- [4] P. Barberá, “How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S.” [Online]. Available: www.pablobarbera.com
- [5] E. Kubin and C. von Sikorski, “The role of (social) media in political polarization: a systematic review,” *Ann Int Commun Assoc*, vol. 45, no. 3, pp. 188–206, 2021, doi: 10.1080/23808985.2021.1976070.
- [6] C. Raymond, “Bridging or Bonding? Measures of Topic Centrality for Online Political Engagement.”
- [7] S. J. Ahmed, “Information Retrieval and Sentimental Analysis with Databricks,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 459–467, Apr. 2021, doi: 10.32628/cseit2172101.
- [8] S. Saif and S. Wazir, “Performance Analysis of Big Data and Cloud Computing Techniques: A Survey,” in *Procedia Computer Science*, 2018, vol. 132, pp. 118–127. doi: 10.1016/j.procs.2018.05.172.
- [9] J. Blair, C. Y. Hsu, L. Qiu, S. H. Huang, T. H. K. Huang, and S. Abdullah, “Using Tweets to Assess Mental Well-being of Essential Workers during the COVID-19 Pandemic,” in *Conference on Human Factors in Computing Systems - Proceedings*, May 2021. doi: 10.1145/3411763.3451612.
- [10] A. Bramson *et al.*, “Understanding Polarization: Meanings, Measures, and Model Evaluation,” 2017. [Online]. Available: <http://www.journals.uchicago.edu/t-and-c>
- [11] F. Marozzo and A. Bessi, “Social Network Analysis and Mining Analyzing Polarization of Social Media Users and News Sites during Political Campaigns.”
- [12] L. Belcastro, R. Cantini, F. Marozzo, D. Talia, and P. Trunfio, “Learning political polarization on social media using neural networks,” *IEEE Access*, vol. 8, pp. 47177–47187, 2020, doi: 10.1109/ACCESS.2020.2978950.
- [13] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer, “Political Polarization on Twitter.” [Online]. Available: www.aaii.org

- [14] A. Tyagi, J. Uyheng, and K. M. Carley, "Affective Polarization in Online Climate Change Discourse on Twitter," Aug. 2020, [Online]. Available: <http://arxiv.org/abs/2008.13051>
- [15] Z. He, N. Mokhberian, A. Câmara, A. Abeliuk, and K. Lerman, "Detecting Polarized Topics Using Partisanship-aware Contextualized Topic Embeddings." [Online]. Available: <https://github.com/ZagHe568/>
- [16] S. H. Kim and H. G. Cho, "User-topic modeling for online community analysis," *Applied Sciences (Switzerland)*, vol. 10, no. 10, May 2020, doi: 10.3390/APP10103388.
- [17] D. Naskar, S. Mokaddem, M. Rebollo, and E. Onaindia, "Sentiment Analysis in Social Networks through Topic Modeling."
- [18] E. Lawrence, J. Sides, and H. Farrell, "Self-segregation or deliberation? blog readership, participation, and polarization in american politics," *Perspectives on Politics*, vol. 8, no. 1, pp. 141–157, Mar. 2010, doi: 10.1017/S1537592709992714.
- [19] K. Mentzer, K. Fallon, J. Prichard, and J. D. Yates, "Measuring and Unpacking Affective Polarization on Twitter: The Role of Party and Gender in the 2018 Senate Races," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.
- [20] T. South, B. Smart, M. Roughan, and L. Mitchell, "Information flow estimation: A study of news on Twitter," *Online Soc Netw Media*, vol. 31, Sep. 2022, doi: 10.1016/j.osnem.2022.100231.
- [21] D. Duque and A. E. Smith, "The esTablishmenT Upside down: a Year of Change in brazil * El 'establishment' dado vuelta: un año de cambio en Brasil."
- [22] D. B. L. Farias, G. Casarões, and D. Magalhães, "Radical Right Populism and the Politics of Cruelty: The Case of COVID-19 in Brazil Under President Bolsonaro," *Global Studies Quarterly*, vol. 2, no. 2, Feb. 2022, doi: 10.1093/isagsq/ksab048.
- [23] C. M. Fernandes, L. A. de Oliveira, M. M. de Campos, and V. B. Gomes, "Political Polarization in the Brazilian Election Campaign for the Presidency of Brazil in 2018: An Analysis of the Social Network Instagram," *Int J Soc Sci Stud*, vol. 8, no. 4, p. 119, Jun. 2020, doi: 10.11114/ijss.v8i4.4837.
- [24] M. L. Layton, A. E. Smith, M. W. Moseley, and M. J. Cohen, "Demographic polarization and the rise of the far right: Brazil's 2018 presidential election," *Research and Politics*, vol. 8, no. 1, 2021, doi: 10.1177/2053168021990204.
- [25] J. Areal, "'Them' without 'us': negative identities and affective polarization in Brazil," *Political Research Exchange*, vol. 4, no. 1, 2022, doi: 10.1080/2474736X.2022.2117635.
- [26] C. Karras, A. Karras, D. Tsohis, K. C. Giotopoulos, and S. Sioutas, "Distributed Gibbs Sampling and LDA Modelling for Large Scale Big Data Management on PySpark," Nov. 2022, pp. 1–8. doi: 10.1109/seed-cccsm57760.2022.9932990.
- [27] S. Vijayarani and M. P. Research Scholar, "Preprocessing Techniques for Text Mining-An Overview."

- [28] V. Kocaman and D. Talby, "Spark NLP: Natural Language Understanding at Scale[Formula presented]," *Software Impacts*, vol. 8, May 2021, doi: 10.1016/j.simpa.2021.100058.
- [29] D. Ramachandran and R. Parvathi, "Analysis of Twitter Specific Preprocessing Technique for Tweets," in *Procedia Computer Science*, 2019, vol. 165, pp. 245–251. doi: 10.1016/j.procs.2020.01.083.
- [30] S. Kumova Metin and B. Karaođlan, "STOP WORD DETECTION AS A BINARY CLASSIFICATION PROBLEM," *ANADOLU UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY A - Applied Sciences and Engineering*, vol. 18, no. 2, pp. 1–1, Jun. 2017, doi: 10.18038/aubtda.322136.
- [31] D. Kumawat and V. Jain, "POS Tagging Approaches: A Comparison," 2015.
- [32] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic N-grams as machine learning features for natural language processing," *Expert Syst Appl*, vol. 41, no. 3, pp. 853–860, 2014, doi: 10.1016/j.eswa.2013.08.015.
- [33] Y. Xu, Y. Yin, and J. Yin, "Tackling topic general words in topic modeling," *Eng Appl Artif Intell*, vol. 62, pp. 124–133, Jun. 2017, doi: 10.1016/j.engappai.2017.04.009.
- [34] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," 2003.
- [35] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Front Artif Intell*, vol. 3, Jul. 2020, doi: 10.3389/frai.2020.00042.
- [36] D. Buenano-Fernandez, M. Gonzalez, D. Gil, and S. Lujan-Mora, "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach," *IEEE Access*, vol. 8, pp. 35318–35330, 2020, doi: 10.1109/ACCESS.2020.2974983.
- [37] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.
- [38] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring Topic Coherence over many models and many topics," Association for Computational Linguistics, 2012. [Online]. Available: <http://mallet.cs.umass.edu/>
- [39] J. Yin and J. Wang, "A Dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 233–242. doi: 10.1145/2623330.2623715.
- [40] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, [Online]. Available: <http://arxiv.org/abs/2203.05794>
- [41] D. A. Schult, "LA-UR-Title: Author(s): EXPLORING NETWORK STRUCTURE FUNCTION USING NETWORKX DANIEL SCHULT PROCEEDINGSITALK SCIPY 08 SWART DYNAMICS, AND NATIONAL LABORATORY Exploring network structure, dynamics, and function using NetworkX."

- [42] F. A. Parand, H. Rahimi, and M. Gorzin, "Combining fuzzy logic and eigenvector centrality measure in social network analysis," *Physica A: Statistical Mechanics and its Applications*, vol. 459, pp. 24–31, Oct. 2016, doi: 10.1016/j.physa.2016.03.079.
- [43] R. J. A. Almeida, "LeIA - Léxico para Inferência Adaptada," *Github*, 2018. <https://github.com/rafjaa/LeIA> (accessed Aug. 27, 2022).
- [44] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," 2014. [Online]. Available: <http://sentic.net/>
- [45] M. Fraundorfer, "Zero hunger for the world-Brazil's global diffusion of its zero hunger strategy," *Austral: Brazilian Journal of Strategy and International Relations*, vol. 2, no. 4, pp. 91–116, 2013, doi: 10.22456/2238-6912.40267.

9. APPENDIX (OPTIONAL)

9.1. SNSCRAPE COMMANDS

Extracting tweets from Jair Bolsonaro:

```
snscape --jsonl --progress --since 2022-01-01 twitter-search "from:jairbolsonaro until:2022-11-01" > bolsonaro/bolsonaro_tweets.json
```

Extracting tweets from Lula:

```
snscape --jsonl --progress --since 2022-01-01 twitter-search "from:LulaOficial until:2022-11-01" > bolsonaro/lulaoficial_tweets.json
```

Extracting tweets from Bolsonaro:

```
snscape --jsonl --progress --since 2022-01-01 twitter-search "@jairbolsonaro until:2022-11-01" > bolsonaro/bolsonaro_mentions.json
```

```
snscape --jsonl --progress --since 2022-01-01 twitter-search "@LulaOficial until:2022-11-01" > bolsonaro/lulaoficial_mentions.json
```

9.2. SPARK NLP PIPELINE

```
def preprocess(df):
    sparknlp.start()
    document = DocumentAssembler().setInputCol("content_cleaned").setOutputCol("document")

    cleanUpPatterns = ["#", "@"]
    documentNormalizer = DocumentNormalizer().setInputCols("document").setOutputCol("normalizedDocument").setAction("clean").setPatterns(cleanUpPatterns)
        .setReplacement(" ") \
        .setPolicy("pretty_all") \
        .setLowercase(True)

    token = Tokenizer().setInputCols(["normalizedDocument"]).setOutputCol("token")

    normalizer = Normalizer().setInputCols(["token"]).setOutputCol("text_processed").setLowercase(True)

    stop_words = StopWordsCleaner.pretrained("stopwords_pt", "pt").setStopWords(nltk_stop_words).setInputCols(["text_processed"]).setOutputCol("cleanTokens")

    lemmatizer = LemmatizerModel.pretrained("lemma", "pt").setInputCols(["cleanTokens"]).setOutputCol('lemmatized')

    pos_tagger = PerceptronModel.pretrained("pos_ud_bosque", "pt").setInputCols(["normalizedDocument", 'cleanTokens']).setOutputCol('pos')

    allowed_tags = ['<ADJ>+<NOUN>', '<NOUN>+<NOUN>']
    chunker = Chunker() \
        .setInputCols(["normalizedDocument", 'pos']) \
        .setOutputCol('ngrams') \
        .setRegexParsers(allowed_tags)

    nGrams = NGramGenerator() \
        .setInputCols("cleanTokens") \
        .setOutputCol("ngrams") \
        .setEnableCumulative(True) \
        .setDelimiter(" ") \
        .setN(2)

    finisher = Finisher() \
        .setInputCols(["ngrams"]) \
        .setOutputCols("ngrams")

    pipeline = Pipeline().setStages([document, documentNormalizer, token, normalizer, stop_words, pos_tagger, nGrams, finisher])
    pipelineModel = pipeline.fit(df)
    result = pipelineModel.transform(df)
    return result.persist()
```

9.3. REPRESENTATIVE DOCUMENTS PER TOPIC

user	topic	content_cleaned
		Encontro ontem em São Paulo com . Conversamos sobre os desafios do país, o combate a fome, sobre o amor como forma de combate a política de ódio, sobre fé e a construção de um Brasil mais fraterno e solidário.
LulaOficial	0	🗨️
LulaOficial	1	Sendo recebido na Bahia por , e. E amanhã estaremos juntos lá na Fonte Nova.
LulaOficial	2	Quem não sabe conviver com a democracia, aposta em fake news! Mentiras sobre urnas, TSE e STF aumentaram 400% em 7 meses. Acompanhe o 🗨️ e vacine-se contra fake news
LulaOficial	3	Quando tomei posse, falei que se quando terminasse meu mandato eu tivesse conseguido levar café da manhã, almoço e janta para a mesa dos brasileiros, eu teria feito a obra da minha vida. E nós conseguimos acabar com a
LulaOficial	4	Todos juntos para restituir o diálogo, o respeito, os direitos dos trabalhadores e a democracia.
LulaOficial	5	Lula participa de entrevista ao vivo para a rádio Liberal FM de Belém do Pará
LulaOficial	6	Eu nunca joguei a culpa das coisas nas costas dos governadores. Não é possível governar o Brasil sem uma boa relação com os governadores e os prefeitos. Não é porque ele foi eleito presidente que pode ficar se achando
LulaOficial	7	Durante o Governo do PT, 82,9% dos acordos dos trabalhadores eram de aumento real de salário, acima da inflação. A vida era melhor e o povo consumia mais. Veja no link: - Teto do financiamento para o curso de medicina aumentou em mais de 20%, em 2022.
jaírbolsonaro	8	- Os cursos da área receberam um incremento de 22,8% e passarão do atual valor, R\$ 42.983,70, para R\$ 52.805,66. O teto do financiamento dos demais cursos permanece no valor de R\$ 42.983,70.
jaírbolsonaro	9	- Avanços no Eixo Norte da transposição do São Francisco. Cenas recentes da barragem de Jati / Ceará. - Vídeo canal do YouTube "Lorim Aventuras".
LulaOficial	10	O Brasil tem artistas excepcionais que aqui são desprezados porque temos um governo que não gosta de cultura. Não existe país democrático sem investimento em cultura. E o governo tem que investir na cultura porque é um
LulaOficial	11	Estive com companheiras e companheiros da União por Moradia Popular. Essa pauta é uma das mais importantes hoje no Brasil. Moradia digna não deveria ser privilégio! Precisamos reconstruir o que esse governo destruiu. - Petrobrás reduz o preço da gasolina: - A partir de amanhã, 20/julho, o preço médio da gasolina para as distribuidoras passa de R\$ 4,06 para R\$ 3,86 por litro.
jaírbolsonaro	12	- A redução é de 5,18%. Brevemente o Brasil terá uma das "gasolina" mais baratas do mundo.
LulaOficial	13	Foi no nosso governo que pela primeira vez uma mulher virou Presidenta da República desse país. Uma mulher que aos 20 anos de idade tinha sido presa, condenada e torturada. E ela virou Presidenta.
LulaOficial	14	Ato Vamos Juntos Pelo Brasil com Lula e Paulo Dantas em Maceió
LulaOficial	15	Atenção Salvador: sábado tem ato com Lula na Fonte Nova! Cadastre-se no site, participe e vamos juntos pela Bahia e pelo Brasil
LulaOficial	16	O Brasil está pronto para retomar o seu protagonismo na luta contra a crise climática, protegendo todos os nossos biomas, sobretudo a Floresta Amazônica. Em nosso governo, fomos capazes de reduzir em 80% o desmatamento na Amazônia. Agora, vamos lutar pelo desmatamento zero.

9.4. NETWORK ANALYSIS

Vertices and Edges

```
1 producers_edges = df_replies.select(F.col("producer").alias("src"),F.col("conversationId").alias("dst")).distinct()
2 consumers_edges = df_replies.select(F.col("user").alias("src"),F.col("conversationId").alias("dst")).distinct()
3 edges = producers_edges.union(consumers_edges)
```

- ▶ producers_edges: pyspark.sql.dataframe.DataFrame = [src: string, dst: long]
- ▶ consumers_edges: pyspark.sql.dataframe.DataFrame = [src: string, dst: long]
- ▶ edges: pyspark.sql.dataframe.DataFrame = [src: string, dst: long]

Command took 0.20 seconds -- by felipesibh@gmail.com at 1/15/2023, 12:39:57 PM on polarization (clone)

Cmd 8

Prepare Data

```
1 df_edges = edges.toPandas()
2 edges_list = df_edges[["src", "dst"]].values.tolist()
```

- ▶ (3) Spark Jobs

Command took 34.44 seconds -- by felipesibh@gmail.com at 1/15/2023, 12:40:11 PM on polarization (clone)

Cmd 9

Create Graph

```
1 G = nx.DiGraph()
2 edges_list = [(item[0],item[1]) for item in edges_list]
3 G.add_edges_from(edges_list)
```

Command took 18.32 seconds -- by felipesibh@gmail.com at 1/15/2023, 12:40:38 PM on polarization (clone)

Cmd 10

Compute eigenvector centrality

```
1 ec = nx.eigenvector_centrality(G,max_iter=100)
```

Command took 9.01 seconds -- by felipesibh@gmail.com at 1/14/2023, 10:53:49 PM on polarization



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa