Semina
Ciências Exatas e Tecnológicas

e-ISSN 1679-0375
STATISTIC

ORIGINAL ARTICLE DOI 10.5433/1679-0375.2025.v46.52599

iThenticate

Topic Modeling of Bot Preferences in Tweets from the COVID-19 Parliamentary Inquiry in Brazil

Modelagem de Tópicos sobre Preferências de Bots em *Tweets* da CPI da COVID-19 no Brasil

Gabriel Thompson Marques Arruda[1], Anderson Castro Soares de Oliveira[2], Lia Hanna Martins Morita[2], José Nilton da Cruz[2]

Received: March 26, 2025 *Received in revised form: June 11, 2025* *Accepted: June 23, 2025* *Available online: July 8, 2025*

ABSTRACT

Twitter is a microblogging social network that allows users to send and receive short messages in text and image format, being one of the most widely used platforms of its kind. Given its relevance in various aspects of contemporary society, including politics, many users operate automated profiles (bots) that post hundreds or even thousands of tweets. This study applies the Latent Dirichlet Allocation (LDA) technique to identify whether bot users show preference for any of the modeled topics. Data were collected via the Twitter API between April 18 and May 30, 2021, using the keywords "CPI" (Parliamentary Inquiry Committee) and "COVID," resulting in 459,145 tweets in Portuguese from 109,027 distinct users. These users were analyzed through the Pegabot platform, which estimates the probability of an account being a bot. After preprocessing, only the users with 100 or more tweets during the period were retained, resulting in 26,966 observations from 189 accounts. LDA identified four main Topics: 1 - Health Secretary; 2 - senator Renan Calheiros; 3 - president Jair Bolsonaro; and 4 - the government. In all topics, bot accounts posts were more frequent than human users posts, with Topic 1 having the lowest proportion of accounts classified as bots.

keywords natural language processing, social media analysis, text mining, automated accounts detection, information dissemination

RESUMO

O Twitter é uma rede social do tipo microblog que permite o envio e recebimento de mensagens curtas em texto e imagem, sendo uma das plataformas mais utilizadas nesse formato. Dada sua relevância em diversos aspectos da sociedade contemporânea, inclusive na política, muitos usuários operam perfis automatizados (*bots*) que disparam centenas ou milhares de tweets. Este trabalho aplica a técnica de Alocação Latente de Dirichlet (LDA) para identificar se há preferência de usuários robôs por determinados tópicos modelados. Os dados foram coletados pela API do Twitter, entre 18 de abril e 30 de maio de 2021, usando as palavras-chave "CPI" e "COVID", totalizando 459.145 tweets em português, de 109.027 usuários distintos. Esses usuários foram analisados pela plataforma Pegabot, que estima a probabilidade de uma conta ser robô. Após o pré-processamento, foram mantidos apenas usuários com 100 ou mais mensagens no período, resultando em 26.966 observações de 189 contas. A aplicação do LDA identificou quatro tópicos principais: 1- Ministério da Saúde; 2 - senador Renan Calheiros; 3 - presidente Jair Bolsonaro; e 4 - governo em geral. Em todos os tópicos, bots publicaram mais que humanos, sendo que o Tópico 1 teve a menor proporção de contas classificadas como robôs.

palavras-chave processamento de linguagem natural, análise de mídias sociais, mineração de texto, detecção de contas automatizadas, disseminação de informações

[1] BSc in Statistics, UFMT, Cuiabá, MT, Brazil. gabrielthompson27@hotmail.com

[2] Prof. Dr., Department of Statistics, UFMT, Cuiabá, MT, Brazil. anderson.oliveira@ufmt.br, lia.morita@ufmt.br, jose.cruz@ufmt.br

Introduction

The internet has revolutionized social interactions, providing users with the ability to create, modify, interpret, and evaluate a vast array of content. In this context, social networks have emerged as the primary platforms for online interaction (Ciribeli & Paiva, 2011). According to the 2023 global report on the digital world, approximately 5 billion people are users of at least one social media platform, representing more than half of the global population (Kemp, 2021). This number indicates that digitalization is a consolidated reality, with a massive volume of data being continuously generated.

Social media platforms have become privileged spaces for the production and dissemination of various types of content, transforming users into active producers of information, opinions, and ideas (Assenmacher et al., 2020). To maximize information dissemination, the activity of automated profiles (bots) on social media has become increasingly common. These bots are programmed to focus on specific content and to attract the attention of targeted segments of users (Assenmacher et al., 2020; Yang et al., 2019).

Given the large volume of data generated daily and the increasing presence of bots, it is essential to develop tools to monitor and analyze the behavior and patterns of these automated agents (Yang et al., 2019). The process of text mining involves multiple stages, including information extraction, data cleaning, classification, and clustering. Topic modeling, particularly through Latent Dirichlet Allocation (LDA), aims to classify text by identifying latent topics within documents. LDA is a probabilistic modeling technique that assumes each document is a mixture of topics, and each topic is characterized by a distribution over words.

The vast amount of data on social media poses challenges to traditional analytical methods. In the present study, the LDA technique was employed for topic modeling, enabling a deeper understanding of user behavior and the operational patterns of bots on social networks. Given the high volume of data generated on Twitter and the need to structure this information, the main objective of this study is to apply topic modeling techniques, with emphasis on the LDA model, to identify the preferences of users previously classified as bots by the Pegabot platform.

The application of the LDA model to Twitter involves collecting, organizing, and modeling data using the aforementioned technique in order to determine whether users identified as bots exhibit preferences for specific topics generated by the model. Thus, the study seeks to answer the following research question: how can the LDA technique be used to identify and analyze the influence of bot activity in the dissemination of information and the formation of topics on Twitter?

Theoretical framework

Social networks

Social networks are essential tools in contemporary society, connecting individuals and organizations based on shared interests, and facilitating real-time communication and information dissemination. In 2023, approximately 5 billion people used social media platforms, representing 62.3% of the global population (Kemp, 2021).

Platforms such as Facebook, Instagram, LinkedIn, Twitter, YouTube, Spotify, and Netflix serve distinct roles, ranging from social interaction and content sharing to business promotion and information diffusion. The broad reach of these media enables the amplification of discourse, the construction of digital identities, and the emergence of new forms of political and cultural engagement (Recuero, 2009).

In addition to fostering social and professional interactions, social media platforms have become strategic spaces for digital marketing, social mobilization, and news circulation. Companies and influencers use these tools to strengthen brands and engage audiences, while governments and institutions explore their potential for public communication and political campaigns (Recuero, 2009).

However, the widespread use of these platforms also brings challenges such as the spread of misinformation, manipulation of debates through bots, and threats to user privacy. Therefore, understanding the dynamics of social networks and their societal impact is essential for ensuring more aware and responsible use (Kemp, 2021).

User behavior and bot influence

Studies of user behavior on Twitter during politically relevant events highlight the growing influence of bots on social networks, emphasizing the need to understand the impact of automated accounts on the dissemination of information and opinions (de Oliveira et al., 2024). Through data collection and analysis, it was observed that the COVID-19 Parliamentary Inquiry Committee (CPI) was a prominent topic on the platform, with a high volume of tweets posted daily.

The presence of bots in these discussions is significant, as these automated agents can influence information diffusion and shape public perception. Detecting bots and understanding their content dissemination strategies are crucial for conducting contextualized analyses of social media discourse. Bots on Twitter—often new accounts with few followers—frequently post intensively to influence trends and debates, underscoring the need for regulation.

Analyses show that bots and humans tend to focus on different types of content, with bots more often targeting institutional topics. Such manipulation may distort public perception and affect democratic integrity, highlighting the importance of transparency and authenticity in digital interactions.

Additionally, it was found that one-third of the fake news about the pandemic in 2020 followed three main strategies: downplaying the threat, accusing certain actors of benefiting from the disease's spread, and promoting unproven treatments (Paganotti, 2021). Corrections issued by fact-checkers received limited user response, ranging from acceptance to outright rejection, suggesting user apathy or resistance to engaging with fact-checkers or automated accounts.

Bot detection

Santos (2020) analyzed the behavior of Twitter profiles from multiple perspectives. However, the Botometer tool used in their methodology exhibited a significant limitation: its accuracy was restricted primarily to English-language content. This constraint negatively impacted the detection of bots in Portuguese, reducing the effectiveness of identifying accounts responsible for spreading harmful or misleading content.

Despite these limitations, the analysis identified patterns of automated behavior, suggesting the presence of disinformation networks operating on the platform. To complement their investigation, the authors employed Gephi software to visualize the structure of interactions between profiles through information transmission graphs. This approach highlighted central users involved in the online discourse surrounding the dismissal of former Health Minister Luiz Henrique Mandetta during the COVID-19 pandemic in Brazil, demonstrating the role of specific profiles in amplifying the debate and propagating particular narratives.

Costa et al. (2021) used the PEGABOT tool to analyze bot activity in propagating hashtags on Twitter, even after the topics ceased to trend. The study compared two hashtags: #BolsonaroDay, which showed signs of bot-driven amplification, and #TowelDay, which had predominantly human participation. Results showed that over 45% of profiles posting under the first hashtag exhibited signs of bot activity, a proportion that increased to 91% among those posting more than 10 tweets with the same hashtag. Furthermore, 6 out of the 10 most frequent retweeters were classified as bots.

In contrast, #TowelDay showed an opposite pattern, with over 81% of profiles identified as human. Among users posting more than 10 tweets with the hashtag, about 50% were human, and 9 out of the 10 most active retweeters were also classified as human. As a next step, the study proposes making the PEGABOT tool publicly accessible and validating its results using additional methodologies and tools to improve bot detection on social media platforms.

Text mining

Text Mining is a process within Knowledge Discovery in Databases (KDD) that employs analytical and extraction techniques to derive meaningful insights from unstructured textual data. It identifies implicit and useful patterns that would be difficult to retrieve through traditional methods (Feldman & Sanger, 2006; Silge & Robinson, 2017; Žižka et al., 2019). The process involves information extraction, data cleaning, classification, and clustering, ultimately transforming textual data into a structured numerical matrix for analysis.

KDD is an interactive process aimed at identifying valid, useful, and interpretable patterns in data (Fayyad et al., 1996). Its main phases include: domain understanding, creation of the target dataset, data cleaning and preprocessing, data transformation, selection of mining methods, exploratory analysis, model and hypothesis definition, pattern extraction, and interpretation of results for knowledge-based decision-making.

Natural Language Processing (NLP) is fundamental to text mining, combining computational techniques with statistical methods to analyze and interpret human language (Chowdhury, 2003). NLP involves four primary stages: morphological analysis, syntactic analysis, semantic analysis, and pragmatic analysis (Bulegon & Moro, 2010).

Most text data sources are incomplete, noisy, and redundant, making preprocessing an essential step. Key techniques include the removal of hashtags, mentions, white spaces, punctuation, numbers, URLs, and stopwords, as well as case normalization. Stemming algorithms are also applied to reduce morphological word variations and extract root forms (Alvares, 2014; Navega, 2002).

Topic modeling

Topic Modeling is essential for organizing and summarizing large text corpora by using machine learning algorithms to identify textual patterns and group words into latent topics (Steyvers & Griffiths, 2007). One of the most widely used methods is Latent Dirichlet Allocation (LDA), proposed by (Blei et al., 2003), which enables the inference of the latent topic structure within a document collection.

From a probabilistic perspective, LDA employs Dirichlet distributions to estimate topic-term and document-topic relationships, allowing each document to be represented as a mixture of topics. The model operates by sampling multinomial variables for each topic and document, as well as selecting topics and terms for each word position within documents. Due to the interdependence between observed and latent variables, inference in LDA is challenging and typically performed using methods such as Gibbs Sampling or Variational Inference (Blei et al., 2003; Griffiths & Steyvers, 2004).

LDA belongs to the class of topic models in which the dependent variable is qualitative (called topic) and is generated from independent variables (text terms). This method assumes that each topic has a set of words (terms) that can define the entire document, and a document can be a mixture of topics (Krestel & Fankhauser, 2010).

In probabilistic notation, let $P(z)$ denote the topic distribution for a specific document d. Each document has a conditional probability distribution $P(w|z)$, where w are the words and z the topics.

Thus, the generalized expression is given by equation(1):

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) \cdot P(z_i = j), \tag{1}$$

where T is the number of topics, $P(w \mid z = j)$ represents the word distribution for topic j, and $P(z)$ is the topic distribution for document d.

The LDA model estimates topic-term distributions using the Dirichlet distribution, whose probability density function is given by:

$$f(z;\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} z_i^{\alpha_i - 1}, \tag{2}$$

where $z = (Z_1, \ldots, Z_K)$ is a K-dimensional variable, where $0 \leq z_i \leq 1$ and $\sum_{i=1}^{K} z_i = 1$; $\alpha = (\alpha_1, \ldots, \alpha_K)$ are the hyperparameters of the distribution; and $B(\alpha)$ is the Beta function, which can be expressed using the Gamma function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}. \tag{3}$$

In the LDA model, the variables ϕ and θ are defined, where ϕ is an n-dimensional variable, with n representing the vocabulary size, and θ is a K-dimensional variable, with K representing the number of topics.

Assuming a document d_j, the LDA model can be executed as follows:

1. Sample K multinomials $\phi_k \sim Dir(\beta)$, one for each topic k;
2. Sample m multinomials $\theta_j \sim Dir(\alpha)$ for the document d_j;
3. For each position i of the words in document d_j:
   - Select a topic $z_{j,i}$ from the distribution θ_j;
   - Select a word $w_{j,i}$ from the distribution $\phi_{z_{j,i}}$.

Considering the observed and latent variables, the joint distribution is:

$$\begin{aligned} p(z, w, \phi, \theta \mid \alpha, \beta) &= \prod_{k=1}^{K} p(\phi_k \mid \beta) \\ &\cdot \prod_{j=1}^{M} p(\theta_j \mid \alpha) \left(\prod_{i=1}^{V} p(z_{j,i} \mid \theta_j) \cdot p(w_{j,i} \mid z_{j,i}, \phi_{z_{j,i}}) \right). \end{aligned} \tag{4}$$

Based on the model in equation (4), it is evident that there is a strong dependency between observed and latent variables. Therefore, the main challenge in LDA is to estimate:

$$p(z, \phi, \theta \mid w, \alpha, \beta), \tag{5}$$

where w are all the observed words in the document collection.

There are several inference methods for LDA; the most commonly used are *Gibbs Sampling* (Griffiths & Steyvers, 2004) and *Variational Inference* (Blei et al., 2003).

Materials and methods

Data collection

The tweet collection was carried out using the R software (RCore Team, 2020) through the `rtweet` package (Kearney, 2019), which allows access to the Twitter API and enables the collection of up to 18,000 tweets every 15 minutes.

Data were collected daily between April 18 and May 30, 2021, using the keywords “CPI” and “COVID,” resulting in a total of approximately 459,000 tweets in Portuguese. These tweets were generated by approximately 109,000 unique users. Each profile was analyzed by the PEGABOT software (Pegabot, 2018) to estimate the probability of being a bot. The higher the percentage assigned by PEGABOT, the higher the likelihood that the profile is automated rather than operated by a real user.

Data selection

From the total of 109,000 users, only those who posted 100 or more tweets during the collection period were considered, ensuring a dataset composed of users with different probabilities of being identified as bots. For each selected user, all tweets collected between April 18 and May 30, 2021, were analyzed.

Topic modeling

The tweets were preprocessed to minimize irrelevant information contained in the posts. The following steps were applied: text normalization to lowercase, removal of hashtags and mentions, deletion of URLs, removal of punctuation and numbers, among other noise. Preprocessing was performed using functions from the `tm` package (Feinerer et al., 2008) in R (RCore Team, 2020).

In topic modeling, determining the number of topics significantly influences model performance. To determine this value, the Cao Juan metrics (Cao et al., 2009) and Deveaud metrics (Deveaud et al., 2014) were used.

The Cao Juan metrics adopts an adaptive method for selecting the best LDA model, based on density. According to this metric, the optimal number of topics is where the average cosine distance between topics reaches its minimum value. On the other hand, the Deveaud metrics performs an LDA analysis over a range of possible topics. For each LDA model, the number of latent concepts is estimated by optimizing the information divergence D between all topic pairs (k_i, k_j). The optimal number of topics is the one that maximizes the divergence D.

To apply the Cao Juan and Deveaud metrics, the `ldatuning` package (Murzintcev, 2020) in R (RCore Team, 2020) was used.

Once the number of topics was defined and the dataset was preprocessed, topic modeling was performed using the LDA technique. For model fitting, the `topicmodels` package (Grün & Hornik, 2011) in R (RCore Team, 2020) was used.

Tweet classification

Based on the probability of profiles being bots, profiles with a probability lower than 0.7 were classified as non-bots, while those with a probability greater than or equal to 0.7 were classified as bots.

With the LDA model fitted, the probability of each tweet belonging to each topic was extracted. Thus, a given tweet was considered to belong to the topic for which it had the highest score (probability). As a result, each tweet was assigned both the bot classification (yes or no) and its most probable topic. Using this information, the distribution of tweets by topic and user type was analyzed.

After classifying the tweets via the fitted LDA model and obtaining the respective probabilities, the proportions of posts made by bot and non-bot users were calculated. The response variable "bot" is categorical (yes or no), and therefore, a multiple comparison of the proportions of bots across topics was performed using a Generalized Linear Model (GLM) with binomial distribution, through the `glm` function available in the `stats` package in R (RCore Team, 2020).

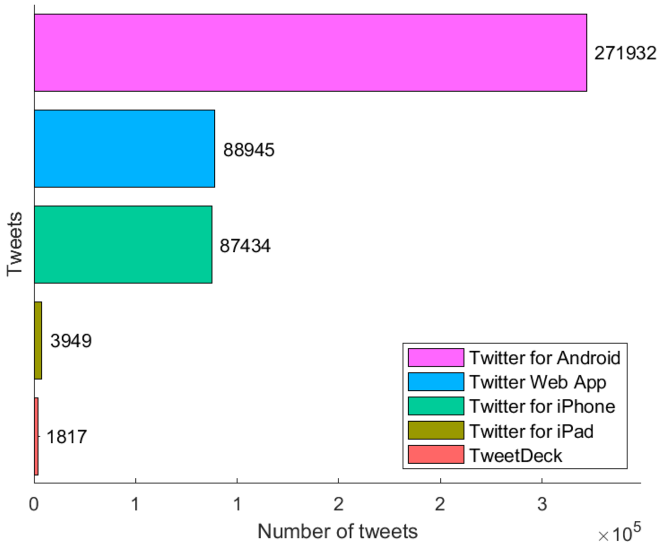
According to Agresti (2018), assuming categorical data follows a binomial distribution is more accurate than using a normal approximation. Subsequently, multiple comparisons were conducted using the `glht` function from the `multcomp` package (Hothorn et al., 2008).

Results and discussion

Database description

During the data collection period, a total of 459,145 tweets were obtained from 109,027 distinct users. These tweets originated from 390 different posting sources, with the five most frequent being *Twitter for Android* (59.22%), *Twitter Web App* (19.37%), *Twitter for iPhone* (19.04%), *Twitter for iPad*, and *TweetDeck* (0.39%), illustrated in Figure 1. The top three sources account for over 97% of the dataset.

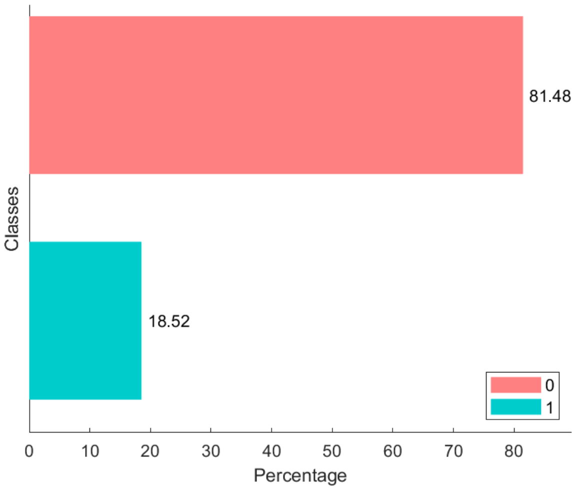
Figure 1 - Number of messages by posting source.



Twitter bot automation can be performed through various custom platforms that enable the programming of specific behaviors, such as automated posting, retweets, likes, and replies to other users. Moreover, these bots may operate using standard clients such as Twitter for Android or iPhone, which are common interfaces for human users. This ability to use standard clients makes it difficult to identify bots based solely on the posting source, as both humans and bots may utilize the same tools to interact with the platform. The use of public and private Twitter APIs also facilitates automation by allowing developers to create scripts and applications that interact automatically with the social network. Therefore, the posting source is not always a reliable indicator of bot activity, since automation can be masked as typical human behavior (Bolsover & Howard, 2019).

Figure 2 shows the classification of users into bots and non-bots, based on the results from the Pegabot platform. A majority of 81.48% of the accounts were identified as non-bots, while 18.52% were classified as bots. This distinction is relevant for understanding how automated accounts may influence the spread of content and engagement patterns within the dataset.

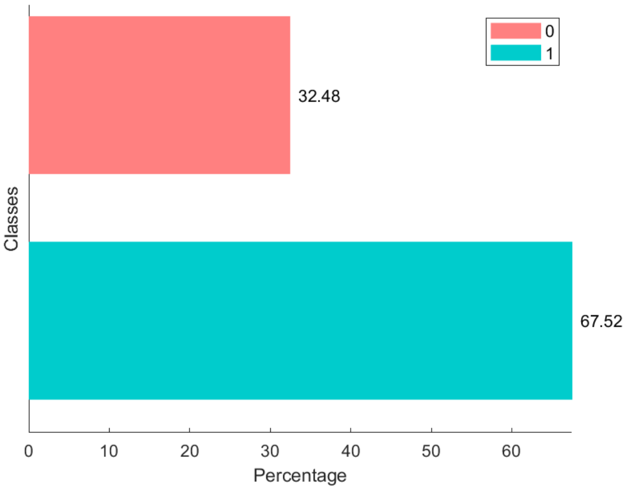
Figure 2 - Percentage of users identified as bots (1) and non-bots (0) in the original dataset.



Liu (2019) conducted a study involving 6,435,932 Twitter accounts and found that 1,084,967 (16.9%) were classified as bots. Similarly, Martini et al. (2021) selected a sample of 122,884 Twitter user accounts that produced 263,821 tweets related to five political discourses in five Western democracies, identifying 27,363 bot accounts (22.0% of all accounts). These studies highlight the significant prevalence of bots on Twitter, particularly in political contexts.

This study focused on users who posted 100 or more tweets, reducing the dataset to 27,120 tweets from 189 distinct accounts. Among these, 75% of users posted up to 156 tweets during the data collection period. User classification revealed that 67.52% were considered bots and 32.48% non-bots, Figure 3, indicating a higher prevalence of bots among the most active users.

Figure 3 - Percentage of users identified as bots (1) and non-bots (0) in the filtered dataset.

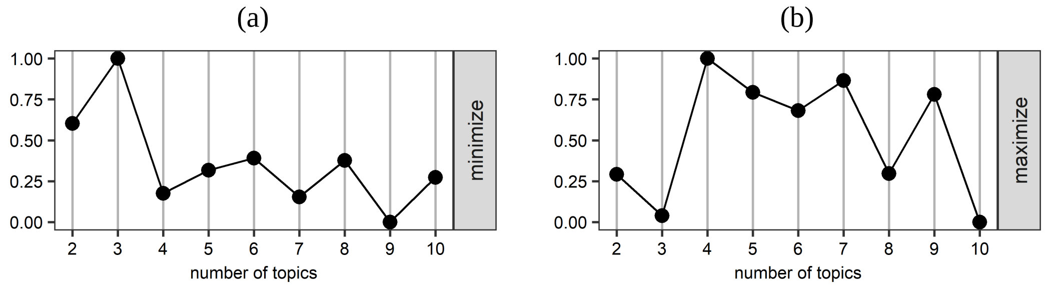


The tendency of bots on Twitter to generate a higher number of posts was also observed in previous studies such as Liu (2019), who emphasized that bots aim to send as many tweets as possible to reach a broader audience. The results of this study support these findings, suggesting that bots play a substantial role in content generation, especially among the most active users. The greater activity of bots, as identified in both our study and in Liu (2019) and Martini et al. (2021), reinforces the importance of considering posting frequency when analyzing bot influence on Twitter.

Latent Dirichlet Allocation

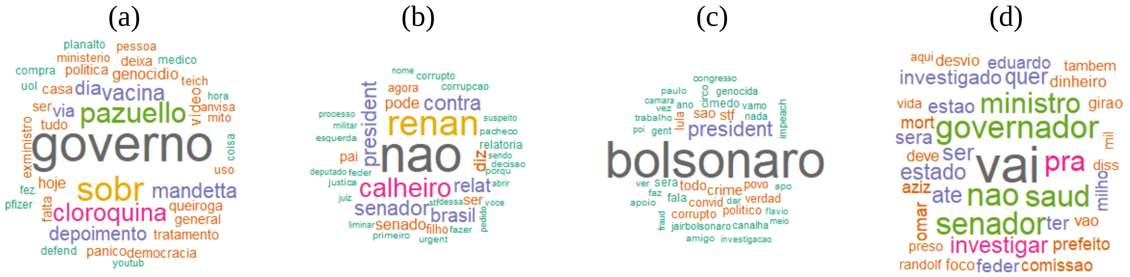
Figure 4 shows the number of topics suggested using the Cao Juan and Deveaud metrics (Cao et al., 2009; Deveaud et al., 2014). According to the Cao Juan metrics, the minimum average cosine distance is achieved with $k = 9$ topics, although a similar value occurs at $k = 4$. For the Deveaud metrics, the maximum information divergence is reached at $k = 4$. Therefore, the LDA model was fitted with 4 topics.

Figure 4 - Suggested number of topics using the metrics: (a) Cau Juan and (b) Deveaud.



After defining the number of topics, the LDA model was fitted using a Gibbs sampler with 50,000 iterations. The word clouds for the four topics generated by the LDA model are presented in Figure 5.

Figure 5 - Word cloud for the 4 generated topics: (a) Topic 1, (b) Topic 2, (c) Topic 3, and (d) Topic 4.



In Figure 5, the most frequent terms for each topic were:

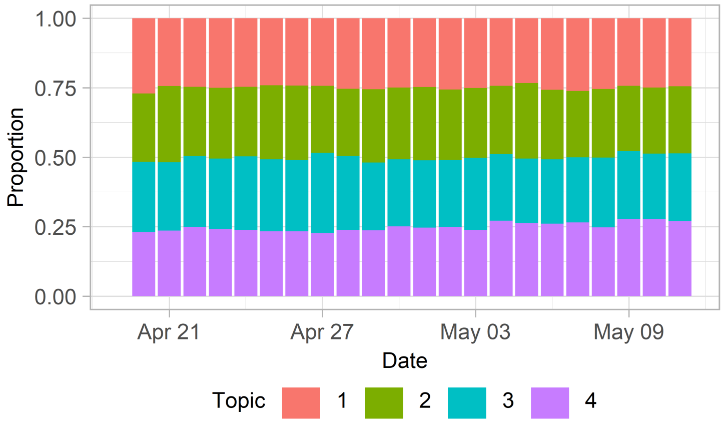
- Topic 1: "governo sobr pazuello cloroquina vacina mandetta depoimento via video hoje genocidio queiroga";
- Topic 2: "nao renan calheiro president senador contra brasil relat diz pode senado ser pai agora filho corrupcao corrupto";
- Topic 3: "bolsonaro president todo crime stf sao covid corrupto povo lula politico verdad medo";
- Topic 4: "vai governador nao senador saud ministro pra investigar ser estado ate quer investigado".

The word clouds help identify the main themes discussed in each topic. Topic 1 is mostly associated with the Health Secretary, Topic 2 refers to Senator Renan Calheiros, Topic 3 focuses on President Bolsonaro, and Topic 4 addresses general issues related to the government.

Tweets are a reflection of ongoing societal discussions, derived from various sources such as newspapers, official websites, and institutional events. Additionally, users themselves act as disseminators of information within the dynamics of the social network (de Sousa, 2015). Therefore, the topics generated tend to reflect the debates that occurred during the data collection period.

Figure 6, which shows the distribution of tweet proportions by topic over time, indicates that all four topics maintained a relatively balanced proportion throughout the data collection period. This suggests that all four topics were widely discussed during this time. Such distribution implies no single topic dominated the discourse, indicating a diversity of interests and concerns among users.

Figure 6 - Distribution of tweet proportions by topic over time.



Furthermore, the temporal analysis of tweets reveals that specific events may have influenced the increase or decrease in discussion of specific topics.

Conclusions

This study presented a detailed analysis of bot activity on *Twitter* during a specific period, employing advanced text mining techniques, such as Latent Dirichlet Allocation (LDA). The focus was on the COVID-19 Parliamentary Inquiry Committee (CPI), with approximately 459,000 tweets in Portuguese collected between April 18 and May 30, 2021, generated by around 109,000 distinct users.

The analysis of tweet distribution by posting source showed that most publications originated from standard Twitter clients, such as *Twitter for Android* and *Twitter Web App*. This highlights the difficulty of identifying bots based solely on the posting source, since both humans and bots may use the same tools to interact on the platform. Therefore, it underscores the importance of employing more sophisticated methods to detect automated behavior, such as analyzing activity patterns and applying machine learning algorithms to identify typical bot characteristics.

The classification of users into bots and non-bots revealed that 18.52% were identified as bots. However, when considering only users who posted more than 100 tweets during the collection period, this proportion increased to 67.52%. This finding is consistent with previous studies, such as those by Liu (2019) and Martini et al. (2021), which also identified a significant presence of bots on *Twitter*. The higher activity levels among bots, as identified in both our study and the referenced literature, reinforce the importance of considering posting frequency when analyzing the influence of bots on social networks.

The application of the LDA technique enabled the identification of four main discussion topics related to the COVID-19 CPI: issues related to the Health Secretary, Senator Renan Calheiros, President Jair Bolsonaro, and the government in general. The balanced distribution of tweet proportions by topic over time indicates that these subjects were widely discussed throughout the data collection period, reflecting the diversity of user interests and concerns.

In all topics, users classified as bots published more tweets than those classified as non-bots. A significant level of bot activity was observed in amplifying information dissemination and influencing public debate. However, in Topic 1, which was associated with the Health Secretary, there was a lower proportion of tweets generated by bots compared to the other topics. These results emphasize the significant influence of bots in spreading information on *Twitter*, especially on topics related to political events and figures.

This study contributes to a better understanding of bot behavior dynamics on *Twitter* and highlights the importance of advanced analytical techniques in identifying automated behavior patterns.

Future research may explore the application of these techniques in other contexts and timeframes, as well as develop more effective methods for detecting and mitigating bot influence on social media platforms. It is essential to continue developing and improving tools for detecting and curbing automated activity on social networks, aiming to preserve the transparency and reliability of information shared online.

Author contributions

G. T. M. Arruda was responsible for conceptualizing, data curation, formal analysis, investigation, methodology, and writing the original draft. **A. C. S. de Oliveira** contributed through supervision, validation, and writing–review and editing. **L. H. M. Morita** also participated in the supervision of the project, as well as in visualization and writing, review, and editing. **J. N. da Cruz** provided technical support, assisted with formal analysis, and contributed to writing, review, and editing.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

The authors would like to thank the Federal University of Mato Grosso (UFMT) and the Núcleo de Estudos, Pesquisas e Inovação em Estatística Aplicada (Center for Studies, Research and Innovation in Applied Statistics) for their support in the development of this study.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Alvares, R. V. (2014). *Algoritmos de Stemming e o Estudo de Proteomas* [Tese de Doutorado]. Universidade Federal do Rio de Janeiro. https://www.pesc.coppe.ufrj.br/uploadfile/1398446767.pdf

Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., & Grimme, C. (2020). Demystifying social bots: On the intelligence of automated social media actors. *Social Media + Society*, *6*(3), 1–14. https://doi.org/10.1177/2056305120939264

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bolsover, G., & Howard, P. (2019). Chinese computational propaganda: Automation, algorithms and the manipulation of information about Chinese politics on Twitter and Weibo. *Information, Communication & Society*, *22*(14), 2063–2080. https://doi.org/10.1080/1369118X.2018.1476576

Bulegon, H., & Moro, C. M. C. (2010). Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. *Journal of Health Informatics*, *2*(2), 51–56. https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/5

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Ciribeli, J. P., & Paiva, V. H. P. (2011). Redes e Mídias Sociais na Internet: Realidades e Perspectivas de um Mundo Conectado. *Mediação*, *13*(12), 57–74. https://revista.fumec.br/index.php/mediacao/article/view/509

Costa, P. H. E. C., Lima, J. R., Marques, R. A., Trindade, D. R., & Komati, K. S. (2021). Estudos de caso de análise de perfis de usuários agrupados por hashtags no Twitter. In Sociedade Brasileira de Computação, *Anais da Escola Regional de Banco de Dados* [Anais]. 16º Escola Regional de Banco de Dados, Santa Maria, Brasil. https://doi.org/10.5753/erbd.2021.17250 de Oliveira, A. C. S., Paixão, C. A., Morita, L. H. M., de Barros, R. C. B., & Ferreira, E. B. (2024). CPI da Covid-19 no Twitter: Uma análise da participação de robôs nas discussões e sentimentos observados. *Esferas*, (29), 1–23. https://doi.org/10.31501/esf.v1i29.14845

de Sousa, M. d. C. E. (2015). A dinâmica da notícia nas redes sociais na internet: A forma de apresentação das postagens no Twitter e no Facebook. *Revista Fronteiras*, *17*(2), 199–212. https://doi.org/10.4013/fem.2015.172.07

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61–84. https://doi.org/10.3166/DN.17.1.61-84

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). (1996). *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54. https://doi.org/10.18637/jss.v025.i05

Feldman, R., & Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30. https://doi.org/10.18637/jss.v040.i13

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Kearney, M. W. (2019). rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, *4*(42), 1829. https://doi.org/10.21105/joss.01829

Kemp, S. (2021). Digital 2021: Global Overview Report. *Datareportal*. https://datareportal.com/reports/digital-2021-global-overview-report

Krestel, R., & Fankhauser, P. (2010). Language Models and Topic Models for Personalizing Tag Recommendation. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.* [Proceedings]. International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, Canadá. https://doi.org/10.1109/WI-IAT.2010.29

Liu, X. (2019). A big data approach to examining social bots on Twitter. *Journal of Services Marketing*, *33*(4), 369–379. https://doi.org/https://doi.org/10.1108/JSM-02-2018-0049

Martini, F., Samula, P., Keller, T. R., & Klinger, U. (2021). Bot, or not? comparing three methods for detecting social bots in five political discourses. *Big Data & Society*, *8*(2), 1–13. https://doi.org/10.1177/20539517211033566

Murzintcev, N. (2020). *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters* [R package version 1.0.2]. https://rdrr.io/cran/ldatuning/

Navega, S. (2002). Princípios Essenciais do Data Mining. Anais do Infoimagem 2002. Cenadem. http://www.intelliwise.com/reports/i2002.pdf

Paganotti, I. (2021). Acolhimento e resistência a correções de fake news na pandemia: a experiência do robô Fátima, da agência Aos Fatos, no Twitter. *Mídia e Cotidiano*, *15*(3), 169–193. https://doi.org/10.22409/rmc.v15i3.47883

Pegabot [Verificador de perfil Twitter]. (2018). https://pegabot.com.br/

RCore Team. (2020). R: A Language and Environment for Statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/

Recuero, R. (2009). Redes sociais na internet (1st ed.). Sulina.

Santos, A. E. G. O. (2020). *Modelo Probabilístico de Tópicos e Estatística Multivariada Aplicados à Análise Textual: Um Módulo de Detecção de Conversas Fora do Contexto para Analisar Conversas em Grupo* [Dissertação de Mestrado.Universidade Federal Rural do Semi-Árido; Universidade do Estado do Rio Grande do Norte]. Repositório. https://ppgcc.ufersa.edu.br/wp-content/uploads/sites/42/2021/02/Disserta%C3%A7%C3%A3oAdriano.pdf

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. O'Reilly Media.

Steyvers, M., & Griffiths, T. (2007). Probabilistic Topic Models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 439–460). Routledge. https://www.routledge.com/Handbook-of-Latent-Semantic-Analysis/Landauer-McNamara-Dennis-Kintsch/p/book/9781138004191

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. https://doi.org/10.1002/hbe2.115

Žižka, J., Dařena, F., & Svoboda, A. (2019). *Text mining with machine learning: principles and techniques*. CRC Press. https://doi.org/10.1201/9780429469275