Exploring Emerging Topics on Antibiotic Use in Brazilian Tweets via Unsupervised Learning

Hudson Mazza[1] **and Lilian Berton**[2]

[1]Universidade da Amazônia (UNAMA)
66060-902– Belém – PA – Brazil

[2]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (UNIFESP)
12247-014– São José dos Campos – SP – Brazil

lberton@unifesp.br, mazzahudson05@outlook.com

Abstract. *Antibiotic misuse plays a critical role in the global spread of antimicrobial resistance, underscoring the need to understand how the public discusses and perceives antibiotic use. This study investigates emerging topics related to antibiotic use among Brazilian Twitter users by applying unsupervised learning techniques. We collected a corpus of Portuguese-language tweets and modeled the textual data using BERTimbau, a language representation model pretrained specifically for Brazilian Portuguese. To uncover latent structures in the data, we applied K-Means clustering on the sentence embeddings, enabling us to identify and interpret thematic groupings within the public discourse. The analysis revealed recurring topics such as respiratory infections, public figures, and self-care dominate discussions. These insights demonstrate the value of combining social media data with unsupervised learning to support public health communication and surveillance strategies in Brazil.*

1. Introduction

Antimicrobial resistance (AMR) has been widely recognized as a pressing global health issue, with numerous studies emphasizing the role of antibiotic misuse in accelerating the spread of resistant pathogens [Organization 2014, Ventola 2015]. Recent research has highlighted the importance of understanding public perceptions and behaviors surrounding antibiotics to inform health policy and communication efforts [McCullough et al. 2016, Roope et al. 2019, Kim et al. 2023]. In this context, social media platforms such as Twitter (actual X) have emerged as valuable data sources for monitoring public discourse and identifying health-related concerns in near real-time [Charles-Smith et al. 2015, Garcia and Berton 2021].

Several studies have leveraged natural language processing (NLP) and machine learning to analyze health conversations on social media. For instance, [Kendra et al. 2015] trained a deep learning classifier on a small set of labeled tweets and enhanced its performance using a larger unlabeled dataset. The classifier achieved 70% accuracy across 9 categories in cross-validation. In the Brazilian context, studies remain limited, although recent efforts have explored the use of BERT-based models, such as BERTimbau [Souza et al. 2020a], to improve the understanding of Portuguese-language social media data.

Additionally, unsupervised learning techniques, particularly clustering algorithms like K-Means, have proven effective in discovering latent themes within unstructured text

data [Aggarwal and Zhai 2012, Jelodar et al. 2019]. When applied to sentence embeddings from transformer-based models, these methods enable more nuanced topic identification compared to traditional techniques such as LDA.

Building on this body of work, our study applies a BERTimbau-KMeans pipeline to explore how Brazilian Twitter users discuss antibiotic use, contributing novel insights into regional patterns of self-medication, misinformation, and public sentiment toward antimicrobial practices.

2. Related work

Previous research investigated social media data concerning bacterial infections. We present here some works related to our study.

[Scanfeld et al. 2010] analyzed 1,000 Twitter posts mentioning "antibiotic(s)" to identify common themes and detect potential misuse or misunderstanding. The tweets were categorized into 11 distinct groups, with major themes including general use, advice, side effects, diagnosis, resistance, and misinformation. A subset of tweets showed evidence of misuse, particularly involving colds, flu, leftover medications, and sharing antibiotics. The findings highlight Twitter's potential as a platform for health communication and surveillance, especially in identifying harmful behaviors or misconceptions. The authors suggest further research to better leverage social media for monitoring and promoting appropriate antibiotic use.

[Andersen et al. 2019] examined yearlong Twitter conversations (Nov 2015–Nov 2016) focused on "antibiotic" and "antimicrobial resistance" (AMR). The analysis revealed that most discussions were concentrated in the U.S. and U.K., with balanced gender participation. Public discourse was largely shaped by retweets and shared links from official sources such as health professionals, media, and government organizations.

[Kim et al. 2023] investigated how AMR content is consumed on Twitter, focusing on engagement patterns with tweets posted by the @AntibioticResis bot, which shares AMR research titles and PubMed links. Using regression models, the researchers analyzed the influence of title wording, pathogen names, and academic visibility on tweet engagement. Results showed that engagement was higher for papers mentioning WHO priority pathogens like Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacteriaceae, as well as for studies with shorter titles. Most followers were healthcare professionals and researchers. The findings indicate that certain pathogens attract more public attention regardless of official priority status, pointing to the need for more tailored public health messaging.

[Zowawi et al. 2015] underscores how platforms like Twitter, Facebook, and YouTube can serve as cost-effective tools for mass health education, citing the Saudi Ministry of Health's use of social media in its MERS-CoV awareness campaign. The authors argue that social media holds strong potential to promote responsible antibiotic use in both the general population and medical communities across the Arabian Peninsula.

[Arquembourg et al. 2025] analyzed over 3,700 French-language social media posts discussing personal experiences with antibiotic ineffectiveness and its impact on quality of life. The results revealed that physical symptoms (78%) and psychological distress (65%) were the most frequently reported impacts. Common themes included

treatment failure, persistent symptoms, side effects, and a lack of clear information. Most posts came from women around 35 years old.

Social media platforms are widely used to monitor discussions about antibiotics, revealing concerns about misuse, self-medication, and therapeutic failures. In Brazil, most research on bacterial resistance focuses on hospital settings [Batista et al. 2023], sanitation reactors [Dropa et al. 2024], and ecological analysis of antibiotic consumption [Boszczowski et al. 2020]. There is still a significant gap in the use of social media as a tool for monitoring or analyzing public discourse related to bacterial resistance and infections.

3. Methodology

This section presents the data collection process, preprocessing, text representation and topic discovery.

3.1. Data Collection

We collected tweets in Brazilian Portuguese containing the keywords related to antibiotic use and disease presented in Table 1.

Table 1. Infection-related and Antibiotic-related Terms

Category	Terms
Infections	infecção urinária, infecção corrente sanguínea, infecção pulmonar, pneumonia, tuberculose, gonorreia, infecção respiratória, tétano, infecções hospitalares, meningite, cólera
Antibiotics	antibiótico, aminoglicosídeo, gentamicina, estreptomicina, amicacina, canamicina, neomicina, plazomicina, quinolona, levofloxacina, gatifloxacina, moxifloxacina, gemifloxacina, polimixina, colistina, cloranfenicol, daptomicina, oritavancina, tigeciclina, eravaciclina, trimetoprima, tetraciclina

To collect the dataset used in this study, we employed an external API service provided by *TwitterAPI.io*. Although this API is not officially affiliated with Twitter, it enables access to publicly available tweets by allowing users to query specific keywords. The service requires authentication via a user-specific API key and supports configurable search parameters such as language and keywords.

We queried tweets in Brazilian Portuguese using our predefined keyword list. All tweets were publicly accessible and collected only from public accounts, ensuring that we did not include any private or protected user data. According to Brazil's General Data Protection Law (LGPD – Law No. 13,709/2018), it is permitted to collect and analyze publicly available data from social media platforms, provided that the processing complies with LGPD principles such as data anonymization, legitimate purpose (in this case, academic research), and non-discrimination. Moreover, the API provider confirmed that this approach complies with all relevant laws and platform policies.

The collection process was automated using the *n8n* workflow automation tool, which executed API queries and stored the results in a structured Google Sheets document. The workflow began with reading the keyword list from a Google Sheets file, then iterating over each keyword individually. For each term, the workflow constructed an

API request specifying the keyword, language filter (Portuguese), and pagination cursor to retrieve tweets in chronological batches without duplication.

The API response for each batch was processed by a custom JavaScript function that extracted and reformatted relevant fields — tweet ID, URL, full text, engagement metrics (likes, retweets, replies, quotes, views), and the creation timestamp in a standardized format. Each processed tweet was appended to a second Google Sheets file serving as the main dataset.

To ensure stability and avoid rate-limit issues, the workflow inserted a two-second delay between requests and maintained internal counters for the number of iterations per keyword, as well as a pagination cursor returned by the API. After each insertion into the dataset, the workflow evaluated whether more tweets were available for the current keyword (based on a *has_next_page* flag). If so, it updated the cursor and repeated the request; otherwise, it advanced to the next keyword in the list. This looping mechanism allowed continuous, paginated scraping while preventing duplicate retrieval.

Each tweet entry included metadata such as tweet text, retweet count, timestamp, and URL. However, only the tweet text was used in the subsequent analysis, as our goal was to apply unsupervised learning techniques (K-Means clustering) on the textual content alone. Figure 1 illustrates the workflow steps in *n8n*, which connects different apps and services using custom logic nodes.

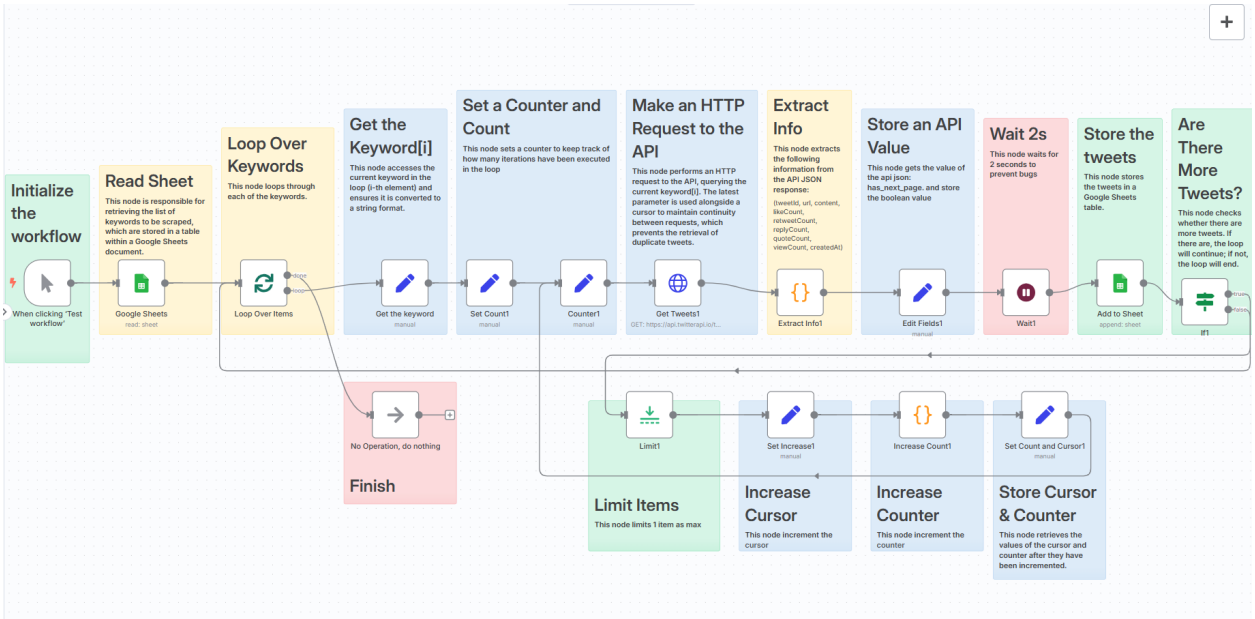

Figure 1. n8n workflow designed for scraping Tweets used in the analysis.

The last data collection was performed on April 29, 2025. For each keyword, the API retrieved tweets in reverse chronological order, starting from the date of the request (April 29, 2025) and going backward in time as far as possible. However, due to high tweet volumes for some popular keywords (e.g., “pneumonia”), or occasional performance limitations of the workflow, the data collection was sometimes interrupted before reaching the full historical depth. As a result, some keywords yielded datasets that go back several years (e.g., to 2018), while others only retrieved more recent tweets. Impor-

tantly, no manual filtering by date was applied, the process depended entirely on how far the API could return results for each term before hitting technical limits.

This variability in historical depth across keywords is a limitation of the dataset and is transparently acknowledged in our methodology. Figure 2 displays the word cloud generated from all collected tweets. It highlights that certain types of infections appear more frequently than others, such as urinary, pulmonary, and respiratory infections, as well as pneumonia, tuberculosis, tetanus, and COVID.

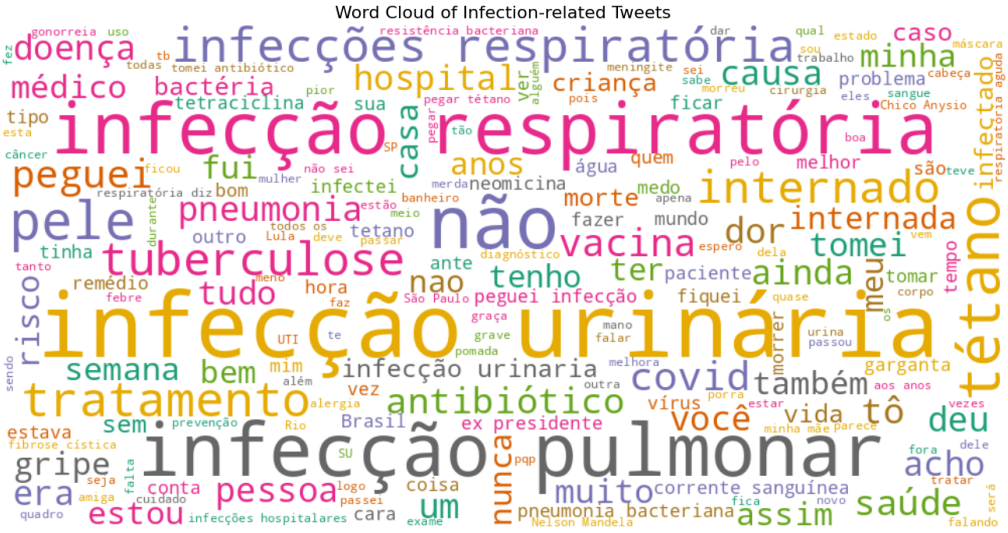

Figure 2. Word cloud of infection-related tweets after preprocessing and stopword filtering.

3.2. Preprocessing

To prepare the dataset for analysis, we implemented a series of preprocessing steps on the tweet texts.

First, all entries were converted to lowercase to ensure case-insensitive comparisons. Then, we removed URLs, user mentions (e.g., "@user"), hashtags, punctuation marks, and non-alphabetic characters using regular expressions. We also eliminated multiple spaces and leading/trailing whitespace to normalize the text format.

After cleaning, we removed stopwords specific to Brazilian Portuguese using a combined list of standard NLTK stopwords and a set of custom-defined tokens, including isolated letters and common informal terms frequently found in social media.

Finally, we applied lemmatization using the spaCy library with its Portuguese language model. Only alphabetic tokens were retained during this step to preserve meaningful words while reducing morphological variation. This preprocessing pipeline helped ensure that the resulting text was semantically consistent and suitable for embedding-based analysis.

3.3. Text Representation

Each tweet was encoded using BERTimbau [Souza et al. 2020b], a transformer-based language model pretrained on large-scale Brazilian Portuguese corpora. We extracted sen-

tence embeddings from the model's output layer to capture nuanced contextual meanings across the dataset.

3.4. Topic Discovery via Unsupervised Learning

To identify thematic clusters of tweets:

- K-Means clustering was applied to the BERTimbau embeddings to group similar tweets into distinct topic clusters.

- The optimal number of clusters was determined using techniques such as the elbow method and silhouette scores.

- Cluster were analyzed to interpret dominant topics, supported by keyword frequency and representative tweet sampling.

3.5. Topic Interpretation and Validation

Each cluster was manually inspected to validate the coherence and relevance of emerging topics. The analysis focused on identifying public concerns, misconceptions, and behaviors related to antibiotic use, such as self-medication, healthcare access, and regulatory issues.

4. Results

Figure 3 presents the relationship between the number of clusters k and the corresponding inertia (sum of squared distances from each point to its assigned cluster center). As expected, the inertia decreases as k increases, since more clusters reduce the distance between data points and their centroids.

The key insight from the elbow method is to identify the point where the rate of decrease in inertia sharply slows down, this forms an "elbow" in the curve. This point is considered a good trade-off between model complexity and performance, as adding more clusters beyond this value yields diminishing returns in reducing inertia.

In the presented graph, the elbow appears around k = 4 or 5, suggesting this is the optimal number of clusters. Choosing k at the elbow helps avoid overfitting and ensures that the clusters are meaningful without being overly granular.

Figure 4 shows the visual representation of the clusters for k = 5, obtained using t-SNE. While some cluster separation is evident, there is still noticeable overlap among most of data points.

To better characterize the discourse in each cluster, we examined the most frequent terms and assigned descriptive thematic labels based on semantic patterns. Table 2 summarizes the main terms and proposed interpretations for each cluster.

These topics reflect a diverse set of narratives circulating on social media. **Clusters 0 and 2** focus on clinical concerns and the severity of respiratory and infectious diseases, such as pulmonary infection, pneumonia, Covid, Tetanus, Tuberculosis. Respiratory infections caused by viruses such as respiratory syncytial virus (RSV), influenza, and coronaviruses are highly prevalent, especially among children living in vulnerable communities. Hospital-acquired pulmonary infections also contribute significantly to patient morbidity, with pneumonia being the most frequent healthcare-associated infection

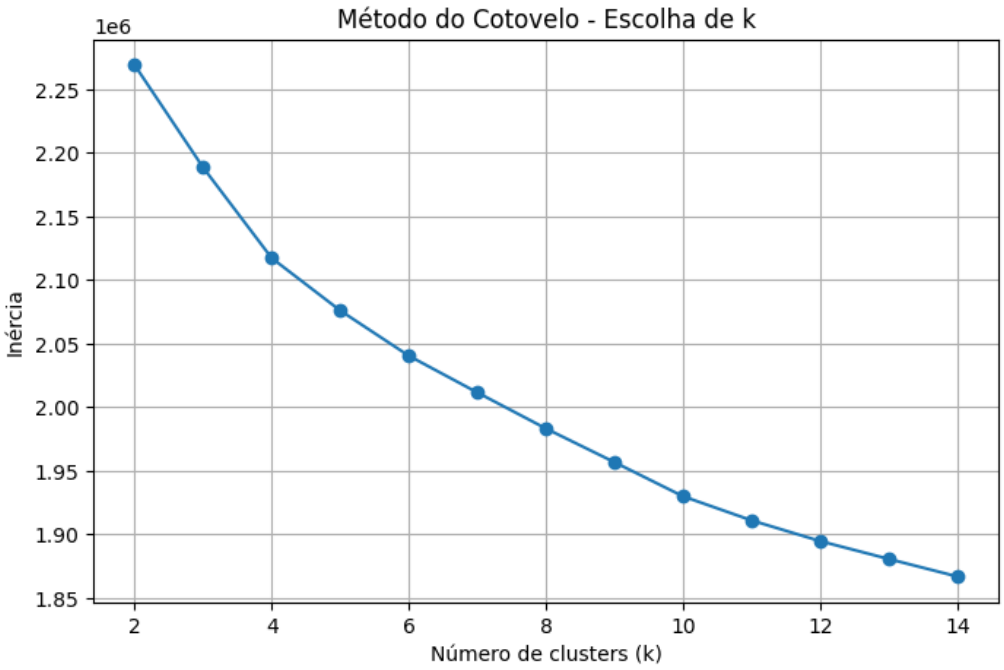

Figure 3. Elbow method indicating the optimal number of clusters for K-Means.

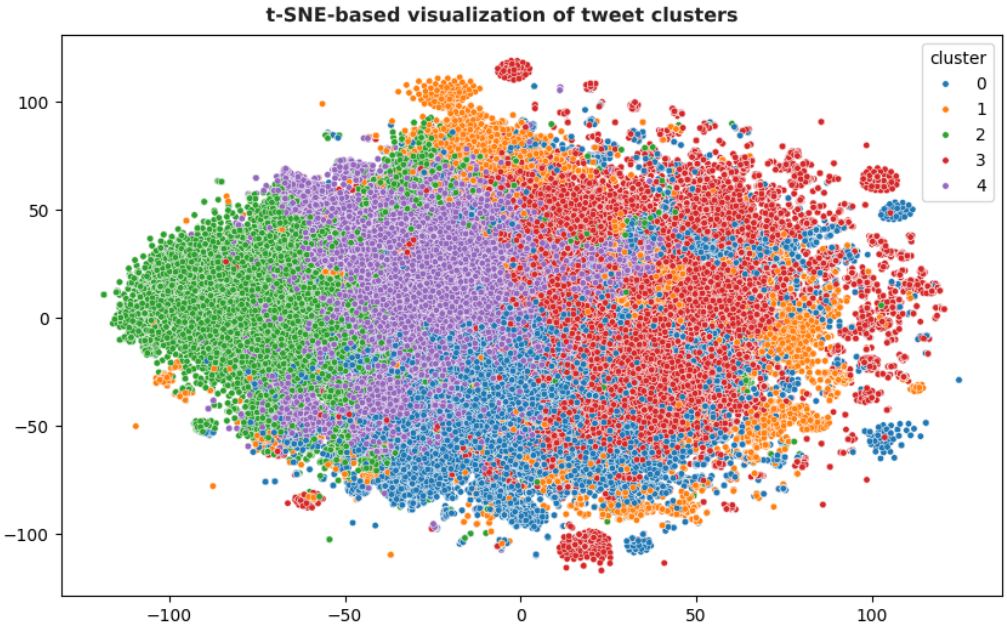

Figure 4. 2D visualization of tweet clusters generated using t-SNE.

in intensive care units. Community-Acquired Pneumonia (CAP) leading cause of hospitalization and death in Brazil. In 2017 alone, CAP caused nearly 600,000 hospital admissions and over 52,000 deaths [Gomes 2018]. Brazil records approximately 69,000 deaths from Tuberculosis between 2005–2010; it remains the second-largest cause of infectious death after HIV/AIDS [Cardoso and Vieira 2016].

Clusters 1 and 3 associate infection-related events with political or public fig-

ures, showing how health discourse can be shaped by social context. Names such as Lula appears, since in 2025 Brazil's President Luiz Inácio Lula da Silva (79) was hospitalized after experiencing vertigo and was diagnosed with labyrinthitis, an inner ear inflammation affecting balance and hearing. In June 2013, former South African President and global icon Nelson Mandela (aged 94) was hospitalized due to a recurrence of pneumonia and related lung complications; this was one in a series of hospitalizations that year. The Venezuelan government reported that Chávez was suffering from a severe respiratory infection and experiencing respiratory decline after surgery in Cuba during 2012 - 2013, with breathing difficulties requiring treatment in a hospital.

Cluster 4 highlights routine health experiences and self-medication practices, reinforcing the importance of monitoring informal public communication on antibiotic use.

Table 2. Topics identified from K-Means clusters based on BERTimbau embeddings

Cluster	Most Frequent Terms	Proposed Topic Name
0	infecção, pulmonar, tratamento, pneumonia, covid, vacina	Respiratory Infections and Medical Treatment
1	infecção urinária, pulmonar, Lula, Mandela, internação	Public Figures and Hospitalization Cases
2	tétano, tuberculose, infectar, morrer, achar	Severe Infectious Diseases and Public Perception
3	infecção respiratória, médico, internação, Chávez	Respiratory Hospitalizations with Political References
4	infecção urinária, sintomas cotidianos, tomar remédio	Everyday Reports and Self-care for Infections

Figure 5 shows the distribution of tweets across different topics. It is evident that tweets concerning respiratory infections and daily reports are the most frequent. This highlights the dual importance of addressing both clinical and social dimensions in public health communication strategies.

5. Discussion

This study offers a novel perspective by revealing that social media discourse about infections is shaped not only by clinical concerns but also by sociopolitical contexts and informal, everyday health experiences. While previous research has extensively examined public health communication on platforms like Twitter in the context of epidemics and vaccine hesitancy (e.g., [Cinelli et al. 2020], [Kouzy et al. 2020]), our findings underscore that infection-related narratives go beyond these scopes. In particular, they are often framed through the lens of political figures and personal, non-clinical accounts of symptoms and treatment practices.

Clusters 1 and 3 in our analysis reveal how health events involving public figures, such as President Lula da Silva's hospitalization for labyrinthitis or Nelson Mandela's recurrent pneumonia, generate significant discourse and shape collective attention toward certain diseases. This confirms that political figures can act as symbolic references in health communication, reinforcing or even distorting public perceptions of medical conditions.

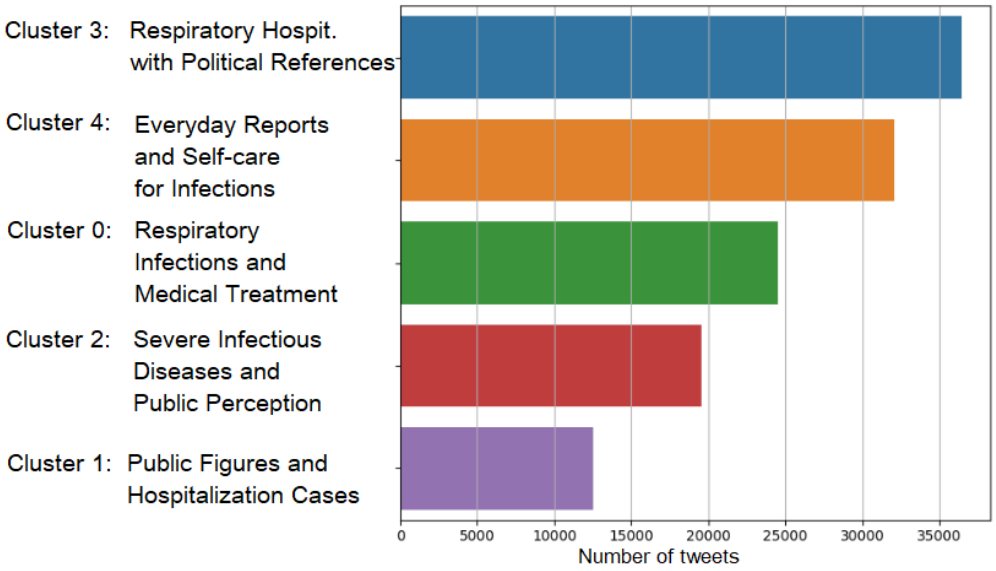

Figure 5. Distribution of tweets by topics.

Simultaneously, Cluster 4 emphasizes how individuals turn to social media to report symptoms and seek or share treatment advice, such as antibiotics or home remedies. This grassroots-level discourse reflects a form of "digital self-care" and underlines the risk of misinformation and self-medication, especially concerning antimicrobial use.

The identification of respiratory diseases, such as pneumonia, tuberculosis, and Covid-19, as central themes (Clusters 0 and 2) aligns with epidemiological data showing their prominence in Brazil's disease burden [Gomes 2018, Cardoso and Vieira 2016]. However, our work goes further by mapping how these concerns manifest in online discourse and interact with social context, offering real-time insights that can complement traditional surveillance methods.

In summary, this study contributes a new insight by showing that public conversations about infections are multi-layered, combining biomedical, political, and experiential narratives. Understanding these dimensions is critical for designing effective and culturally sensitive health communication strategies in the digital age.

6. Conclusions

This study provides a detailed characterization of discourse around infections on social media by clustering tweets using BERTimbau embeddings and K-Means clustering. The thematic analysis of the most frequent terms within each cluster reveals diverse narratives that reflect both clinical realities and sociopolitical contexts.

Overall, this work demonstrates the value of combining advanced NLP techniques with thematic clustering to capture the multifaceted nature of infection-related discourse on social media. Future research could extend these findings by incorporating temporal analysis to track shifts in public attention or by integrating sentiment analysis to assess the emotional tone associated with these topics.

7. Ackowledgement

We thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant 2021/10599-3.

References

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Andersen, B., Hair, L., Groshek, J., Krishna, A., and Walker, D. (2019). Understanding and diagnosing antimicrobial resistance on social media: a yearlong overview of data and analytics. *Health communication*, 34(2):248–258.

Arquembourg, J., Glaser, P., Roblot, F., Metzler, I., Gallant-Dewavrin, M., Nanguem, H. F., Mebarki, A., Voillot, P., and Schück, S. (2025). Discussions of antibiotic resistance on social media platforms: Text mining and mixed methods content analysis study. *JMIR Formative Research*, 9:e37160.

Batista, M. P. B., Cavalcante, F. S., Alves Cassini, S. T., and Pinto Schuenck, R. (2023). Diversity of bacteria carrying antibiotic resistance genes in hospital raw sewage in southeastern brazil. *Water Science & Technology*, 87(1):239–250.

Boszczowski, Í., Neto, F. C., Blangiardo, M., Baquero, O. S., Madalosso, G., de Assis, D. B., Olitta, T., and Levin, A. S. (2020). Total antibiotic use in a state-wide area and resistance patterns in brazilian hospitals: an ecologic study. *The Brazilian Journal of Infectious Diseases*, 24(6):479–488.

Cardoso, T. A. d. O. and Vieira, D. N. (2016). Study of mortality from infectious diseases in brazil from 2005 to 2010: risks involved in handling corpses. *Ciência & Saúde Coletiva*, 21:485–496.

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H. Y., Olsen, J. M., Pavlin, J. A., Shigematsu, M., Streichert, L. C., Suda, K. J., et al. (2015). Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PLOS ONE*, 10(10):e0139701.

Cinelli, M. et al. (2020). The covid-19 social media infodemic. *Nature Human Behaviour*, 4(10):1285–1293.

Dropa, M., da Silva, J. S. B., Andrade, A. F. C., Nakasone, D. H., Cunha, M. P. V., Ribeiro, G., de Araújo, R. S., Brandão, C. J., Ghiglione, B., Lincopan, N., et al. (2024). Spread and persistence of antimicrobial resistance genes in wastewater from human and animal sources in são paulo, brazil. *Tropical Medicine & International Health*, 29(5):424–433.

Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. *Applied soft computing*, 101:107057.

Gomes, M. (2018). Community-acquired pneumonia: challenges of the situation in brazil.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, L., and Lv, Y. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Kendra, R. L., Karki, S., Eickholt, J. L., and Gandy, L. (2015). Characterizing the discussion of antibiotics in the twittersphere: What is the bigger picture? *Journal of medical Internet research*, 17(6):e154.

Kim, H., Proctor, C. R., Walker, D., and McCarthy, R. R. (2023). Understanding the consumption of antimicrobial resistance–related content on social media: Twitter analysis. *Journal of Medical Internet Research*, 25:e42363.

Kouzy, R. et al. (2020). Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *American Journal of Preventive Medicine*, 59(2):261–263.

McCullough, A. R., Parekh, S., Rathbone, J., Del Mar, C. B., and Hoffmann, T. C. (2016). A systematic review of the public's knowledge and beliefs about antibiotic resistance. *Journal of Antimicrobial Chemotherapy*, 71(1):27–33.

Organization, W. H. (2014). Antimicrobial resistance: global report on surveillance.

Roope, L. S. J., Smith, R. D., Pouwels, K. B., Buchanan, J., Abel, L., Eibich, P., El Khoury, A. C., Walker, A. S., and Robotham, J. V. (2019). The challenge of antimicrobial resistance: what economics can contribute. *Science*, 364(6435):eaau4679.

Scanfeld, D., Scanfeld, V., and Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 38(3):182–188.

Souza, F., Nogueira, R., and Lotufo, R. (2020a). Bertimbau: Pretrained bert models for brazilian portuguese. *Portuguese Conference on Artificial Intelligence*, pages 403–417.

Souza, L. F. P., Abreu, V., Cruz, S. J. R., and Pardo, T. A. S. (2020b). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–408. IEEE.

Ventola, C. (2015). The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and Therapeutics*, 40(4):277.

Zowawi, H. M., Abedalthagafi, M., Mar, F. A., Almalki, T., Kutbi, A. H., Harris-Brown, T., Harbarth, S., Balkhy, H. H., Paterson, D. L., and Hasanain, R. A. (2015). The potential role of social media platforms in community awareness of antibiotic use in the gulf cooperation council states: luxury or necessity? *Journal of Medical Internet Research*, 17(10):e233.