Topic Modeling in Brazilian Portuguese Documents on Antimicrobial Resistance

Enrique Reis Susin and Lilian Berton
Instituto de Ciência e Tecnologia - Universidade Federal de São Paulo
Unidade Parque Tecnológico - Avenida Cesare Mansueto Giulio Lattes, n° 1201
Eugênio de Mello, CEP: 12247-014
enrique.susin@unifesp.br, lberton@unifesp.br

Abstract

This study analyzes texts from multiple sources, including social media and news portals, to observe how different sectors of Brazilian society discuss the antimicrobial resistance. The main goal is to support epidemiological surveillance and public policy decisions through computational tools. Three datasets were used: tweets collected between 2008 and 2025 (64,225 documents), news articles from G1 (4,363 documents), and official government publications (.gov.br, 1,515 documents). These sources enable comparative analysis between informal discourse (social media) and institutional or journalistic discourse (official and media outlets). The study applies and compares topic modeling techniques, particularly those designed for Short Text Topic Modeling (STTM), such as GSDMM and BERTopic, to identify discursive trends, semantic patterns, and emerging topics related to antimicrobial resistance. By exploring these distinct contexts, this work demonstrates the potential of Natural Language Processing (NLP) and AI methods as instruments for integrated analysis of public health data in both informal and formal environments.

1 Introduction

Bacterial resistance to antibiotics represents a growing global threat, recognized by the World Health Organization (WHO) as one of the main challenges to sustainable development and global health (World Health Organization, 2024). The increasing misuse and overuse of antimicrobial agents accelerate the emergence of resistant strains, jeopardizing decades of medical progress (Murray et al., 2022; Laxminarayan et al., 2013). It is estimated that, without effective intervention, antimicrobial resistance could cause more deaths than cancer by 2050 (O'Neill, 2016).

This research also aligns with the Sustainable Development Goal (SDG) of the United Nations, especially goal 3 (United Nations General Assembly, 2015), which aims to ensure healthy lives and promote well-being for all. The study focuses on computational approaches that can extract meaningful information from heterogeneous text data sources, providing situational awareness that can aid both scientific and policy decision-making.

The project proposes a large-scale textual analysis using Short Text Topic Modeling (STTM) (Qiang et al., 2020) techniques to detect emerging topics and public discourse patterns related to antimicrobial resistance. It aims to understand how formal and informal communication channels — such as news media, government publications, and social networks — shape social understanding and engagement with this issue. STTM overcomes the severe data sparsity inherent in brief documents by leveraging global context or external knowledge to extract coherent latent themes where traditional algorithms like LDA fail.

We collected three datasets that encompass different linguistic and communicative contexts: informal discussions from Twitter, journalistic narratives from G1, and institutional statements from official government websites. This multi-domain approach allows comparisons across different discursive registers, identifying both overlaps and divergences in how antimicrobial resistance is represented. Ultimately, this work contributes to understanding the discursive and informational dynamics surrounding antibiotic resistance, while demonstrating how AI-based text mining can complement traditional epidemiological methods.

2 Related Work

The challenge of topic modeling in short texts has been widely discussed in the literature, as such texts often lack sufficient context and word co-occurrence for traditional probabilistic models. An analysis of 189 articles by Laureate et al.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 103–110

(2023) concludes that many researchers still misapply LDA, reinforcing the necessity for specialized methods such as STTM. Several approaches have been proposed to address these limitations. This section presents some representative studies.

According to the taxonomy proposed by Qiang et al. (2020), existing works in STTM are divided into three main categories: Dirichlet Multinomial Mixture based methods (e.g., GSDMM), global word co-occurrence based methods (e.g., BTM), and self-aggregation based methods. Representative works illustrate the distinct strategies of each category. For Dirichlet Multinomial Mixtures, GSDMM explorad by Yin and Wang (2014) assumes that each short text is generated by a single topic, a simplification designed to address data sparsity. In the category of global word co-occurrence, BTM used by Yan et al. (2013a) shifts focus from modeling documents to modeling the generation of word pairs (biterms) across the entire corpus, thereby capturing global correlations. Finally, regarding self-aggregation, Quan et al. (2015) proposes aggregating short texts into long pseudo-documents dynamically during the inference process, without relying on external metadata such as authors or hashtags. For a comprehensive list of studies and a detailed taxonomy, we refer the reader to the survey by Qiang et al. (2020).

While Qiang et al. (2020) established the classical taxonomy, recent surveys like Wu et al. (2024) highlight the shift towards Neural Topic Models (NTMs) and Large Language Models (LLMs) to address data sparsity through contextual embeddings and generative capabilities.

STTM in Portuguese were addressed by some works, like Júnior et al. (2022), that investigate how pre-processing affects topic modeling algorithms for short texts (Twitter and Reddit) in Brazilian Portuguese, proposing pipelines that improve performance compared to English-centric approaches. Silva et al. (2021) applies topic modeling to short citizen comments regarding legislative bills, employing embedding-based methods such as BERTopic, which represents the current state-of-the-art for STTM.

To the best of our knowledge, a gap remains in the literature regarding the analysis of antimicrobial resistance through the lens of heterogeneous sources, such as news articles and social media posts. Moreover, no prior research has leveraged Short Text Topic Modeling (STTM) techniques to disentangle latent narratives in this context, specifically differentiating themes like the 'indiscriminate use of antibiotics'.

3 Methodology

This work aims to apply and compare specialized topic modeling approaches for STTM across multiple textual sources related to antimicrobial resistance and antibiotic usage in Brazilian Portuguese.

3.1 Datasets

Three main datasets were collected, each representing a distinct communicative domain:

1. **Tweets (X/Twitter)** — Collected between 2008 and 2025 using keywords related to "bacteria," "antibiotic," and "resistance." This dataset represents spontaneous, fragmented user discourse characterized by informality, repetition, and textual noise (X Corp., 2025). It reflects public perception and daily discourse surrounding bacterial resistance.

2. **Governmental Publications (.gov.br)** — Comprising official statements, health campaigns, and institutional reports from Brazilian government websites (Brazilian Government, 2023). These texts exhibit a formal and technical register, typically related to health policies and official public health communication.

3. **G1 News Articles** — Collected via automated web scraping from the G1 portal's science and health sections (Grupo Globo, 2023). This dataset represents a hybrid discourse that mediates between technical and popular language, translating scientific and institutional content for a broader audience.

Combining these three domains offers a multilayered view of public discourse, enabling comparison of how antimicrobial resistance is treated across informal, journalistic, and institutional contexts. The analysis aims to identify emerging subtopics and discursive patterns linked to antibiotic usage, misinformation, and public perception.

3.2 Data Collection and Preprocessing

Data collection was automated through `TwitterAPI.io`, which provides simplified access to Twitter's API and enables large-scale text retrieval. In accordance with ethical research practices, tweets will not be published or disclosed

in their original form. Only the textual content was analyzed for academic purposes, and no information that could identify individual users was retained or reported. The study strictly adhered to principles of privacy, anonymity, and responsible data use, ensuring that the research complies with institutional and academic ethical standards.

Given the unstructured and noisy nature of tweets, a rigorous preprocessing pipeline was implemented. First, duplicate entries were removed. Next, since terms such as "resistência" and "bactéria" also occur in other languages (e.g., Spanish), automatic language detection was performed using the `LangDetect` library, which applies probabilistic n-gram models to infer language. Only tweets classified as Portuguese ("pt") were retained in the dataset.

Subsequently, textual cleaning steps were applied: removal of user mentions (@user), URLs, hashtags, media indicators, emojis, and onomatopoeic expressions such as "kkkkk" and "rsrsrs." All text was lowercased, extra spaces were reduced, and numeric characters were removed to standardize data for NLP processing.

Tokenization was performed using the Portuguese language model from `spaCy`. Common stop words (e.g., "pra," "de," "da") were removed, and lemmatization was applied to reduce words to their root forms. The entire pipeline was optimized for Portuguese, improving both linguistic and semantic accuracy.

The final preprocessed data were stored in a CSV file containing cleaned, tokenized, and lemmatized tweets. This dataset served as input for topic modeling algorithms described in the following sections.

For the G1 and .gov.br datasets, preprocessing followed a similar structure, differing primarily in data acquisition — automated web scraping from health and science sections. Duplicate titles and repeated news fragments were removed to ensure uniqueness, since government sites frequently republish the same official statements.

The datasets have been fully anonymized and were used exclusively for academic purposes, with no public disclosure of the raw data.

3.3 Keyword Selection

Selecting search terms was a crucial step to ensure thematic coverage and epidemiological relevance. The same list of keywords was used across all three datasets to enable direct comparison. These queries included disease names, drug names, and general antimicrobial resistance concepts.

Table 1 details the selected keywords and the corresponding volume of documents retrieved from each source. In total, 70,103 documents were analyzed: 64,225 tweets, 4,363 G1 articles, and 1,515 government publications. All texts are in Portuguese and discuss topics related to antimicrobial resistance, infections, and antibiotic use.

3.4 Topic Modeling

Two primary algorithms were implemented: **GSDMM** and **BERTopic**. These were chosen because traditional methods like Latent Dirichlet Allocation (LDA) perform poorly in short-text contexts where term co-occurrence is sparse. GSDMM serves as a robust probabilistic baseline, while BERTopic represents the state-of-the-art (SOTA). By contrasting these two models, we aim to capture the methodological evolution over nearly a decade and provide a balanced perspective without diluting the analysis across too many techniques.

The Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) model (Yin and Wang, 2014) assume that each document belongs to a single dominant topic. It groups documents by shared word distributions using Gibbs sampling. This makes GSDMM computationally efficient and suitable for noisy, short, informal texts like tweets.

The algorithm assumes a fixed number of topics K and iteratively reallocates each document based on posterior probability, considering both (1) the number of documents in each topic and (2) the similarity between document terms and topic vocabulary, modulated by Dirichlet priors α and β. This produces compact and interpretable topic clusters.

In contrast, BERTopic (Grootendorst, 2022) leverages dense semantic embeddings combined with dimensionality reduction and density-based clustering. This enables effective short-text topic discovery by using contextual representations from transformer-based models.

We utilized the BERTimbau Base model (Souza et al., 2020), a Transformer pretrained on Brazilian Portuguese, implemented via the `sentence-transformers` library (Reimers and Gurevych, 2019) to generate dense vector representations.

Each tweet was converted into a high-dimensional vector capturing contextual meaning. Clustering was performed using HDBSCAN

Source	Antibiotic	Bacteria	Tuberculosis	UTI	Sepsis	Pneumonia	Amoxicillin	Resistance Variants
GOV	129	181	792	66	13	304	6	24
G1	596	598	594	593	586	597	203	596
Twitter	16,203	—	7,888	6,657	—	5,117	15,871	12,489
Total	16,928	779	9,274	7,316	599	6,018	16,080	13,109

Table 1: Number of documents retrieved per keyword and data source.

(McInnes et al., 2017), a hierarchical density-based algorithm that automatically identifies clusters of varying density and isolates noise points. HDBSCAN proceeds in three main steps: (1) computation of core distances to define local density, (2) construction of a minimum spanning tree based on mutual reachability distance, and (3) extraction of persistent clusters based on stability metrics. This flexibility allows better handling of noisy text data, typical in social media corpora.

4 Experimental Analysis

4.1 Algorithm Configuration and Environment

All experiments were implemented in Python 3.10 using standard NLP and topic modeling libraries. The following packages were employed:

- `transformers` and `sentence-transformers` for embedding generation with BERTimbau Base;
- `hdbscan` for density-based clustering;
- `bertopic` for integrated topic modeling based on embeddings;
- `spaCy` and `langdetect` for text preprocessing;
- `scikit-learn`, `pandas`, and `numpy` for data manipulation and analysis;
- `gsdmm` for implementation of the GSDMM algorithm adapted for short texts.

Multiple parameter combinations were tested, and the most coherent and interpretable topic distributions were selected. Final configurations:

- **GSDMM:** $K = 10$, $\alpha = 0.1$, $\beta = 0.1$, $n\_iters = 30$;
- **BERTopic:**
  - $n\_neighbors = 10$,
  - $n\_components = 5$,
  - $min\_cluster\_size = 7$,
  - $min\_samples = 5$,
  - $metric = \textit{euclidean}$,
  - $cluster\_selection\_method = \textit{eom}$,
  - $diversity = 0.1$.

4.2 Evaluation Criteria

Since topic modeling is an unsupervised task, we used two intrinsic measures widely adopted in topic modeling: topic coherence and topic diversity.

Topic Coherence (C_v) Topic coherence measures the semantic similarity between the most representative words of each topic. Coherent topics are expected to contain words that co-occur frequently in real contexts. We use the C_v metric, which combines word co-occurrence statistics with semantic similarity derived from context windows.

Coherence is computed from the top-N terms of each topic and the preprocessed corpus. Typical values range from 0.3 to 0.7; higher values indicate greater semantic consistency. However, excessively high values can suggest topic overlap and reduced distinctiveness.

Topic Diversity Topic diversity quantifies how distinct topics are from each other, based on the proportion of unique words among the top-N topic terms. A high diversity value implies minimal word overlap between topics, suggesting coverage of multiple distinct themes.

Formally:

$$Diversity = \frac{\text{Number of unique words in topics}}{\text{Total number of words in topics}}$$

Typical diversity values range from 0.3 to 0.9 depending on granularity. Low diversity suggests redundancy; excessively high diversity may indicate incoherent clustering.

4.3 Results and Discussion

The algorithms displayed divergent performance characteristics when processing short and noisy

text data, such as tweets. Table 2 presents the comparative results regarding coherence and diversity metrics.

Algorithm	Coherence (C_v)	Diversity
GSDMM	0.4471	0.7467
BERTopic	0.5032	0.9278

Table 2: Comparison of topic modeling performance.

Although both models yielded relatively low coherence (expected for short and noisy texts), BERTopic achieved superior results in both metrics. Higher coherence indicates more semantically consistent topics, while higher diversity reflects less term overlap among topics.

In practice, BERTopic produced more meaningful clusters, distinguishing relevant thematic groups such as discussions about veterinary antibiotic use, bacterial species, and non-informative content. This differentiation between informative and irrelevant topics is crucial for epidemiological and social research applications.

GSDMM, despite being well-suited for short texts, was less effective on this dataset. Even after parameter tuning, its topics were often redundant or lacked internal coherence. Useful clusters were fewer and more sensitive to random initialization, indicating reduced robustness.

These findings suggest that embedding-based approaches like BERTopic tend to generate richer and more useful representations for heterogeneous short-text corpora compared to frequency-based probabilistic models.

4.4 Topic Analysis

4.4.1 Analysis of Topic Coherence for BERTopic

Figure 1 displays the hierarchical importance of keywords for four distinct topics extracted by the BERTopic model. The visualization relies on Class-based TF-IDF scores to identify the most representative terms. The qualitative analysis of these clusters reveals the model's capability to disentangle broad public health concepts from specific clinical narratives.

- **Topic 0** (General Microbiology and Public Health): This topic aggregates high-level terms such as *mutação* (mutation), *vírus* (virus), *bactéria* (bacteria), and *vacina* (vaccine). The co-occurrence of viral and bacterial terminology alongside "antibiotic" suggests a conflation of pathogen types in public discourse. This is particularly relevant to Antimicrobial Resistance (AMR) surveillance, as it may indicate public confusion regarding the appropriate use of antibiotics for viral infections.

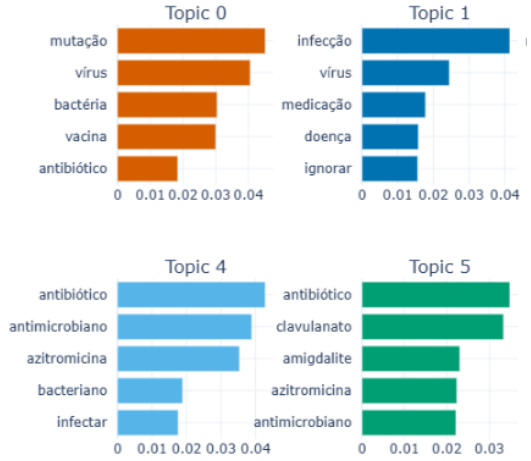

Figure 1: Topic distribution generated by BERTopic.

- **Topic 1** (Behavioral Aspects of Infection): This cluster focuses on the experience of illness, highlighted by terms like *infecção* (infection), *medicação* (medication), and *doença* (disease). Notably, the presence of the verb *ignorar* (to ignore) suggests narratives involving the negligence of symptoms or medical advice, a behavioral pattern often linked to delayed treatment or improper self-medication.

- **Topic 4** (Pharmacological Terminology): In contrast to the more general topics, Topic 4 demonstrates high technical coherence. It groups accurate terminology such as *antimicrobiano* (antimicrobial) and *bacteriano* (bacterial). The prominence of *azitromicina* (azithromycin) reflects its dominance in the corpus, likely driven by its widespread prescription or its visibility in recent health controversies.

- **Topic 5** (Specific Clinical Contexts): This topic provides the most granular insight into patient reporting. It successfully links a specific medical condition, *amigdalite* (tonsillitis), with its standard therapeutic regimens, *clavulanato* (clavulanate) and *azitromicina*. This suggests that BERTopic successfully captured personal health accounts or specific

treatment discussions often found in social media environments.

Overall, the results indicate that the embedding-based approach of BERTopic effectively captures semantic nuances, distinguishing between chemically specific discussions (Topic 5) and conceptual confusion regarding pathogen types (Topic 0).

4.4.2 Exploratory Analysis of Bacterial Resistance and Tuberculosis Narratives

We conducted a thematic analysis to uncover semantic correlations and discourse patterns, assessing their potential utility in public health surveillance. To this end, two primary keyword-driven case studies were established: “Bacterial Resistance” and “Tuberculosis".

Bacterial Resistance. For the “bacterial resistance” keyword, BERTopic was applied to both Twitter and G1 datasets. Twitter data required multiple post-processing steps to remove noisy topics and merge overlapping clusters. G1 articles, being more formal and structured, yielded cleaner topics centered on scientific and health-reporting contexts.

On Twitter, two major discourse axes emerged. The first concerned veterinary and agricultural contexts — antibiotics used in livestock and agricultural production — highlighting the relevance of zoonotic surveillance and antimicrobial control in food systems. The second involved personal experiences and misinformation, featuring colloquial and ironic language around “medicine,” “cure,” and “nonsense,” suggesting both personal anecdotes and misinformation regarding antibiotic use and vaccination.

Refined clustering further revealed two subgroups: one technical (terms such as “plasmid,” “strain,” “KPC,” and “superbacteria”), representing specialized or scientific discussions; and another humorous/offensive group identified as noise and discarded. This iterative refinement underscored the importance of denoising and re-clustering when analyzing informal texts.

In the G1 dataset, the most coherent topics corresponded to hospital outbreaks and epidemiological events (keywords like “hospital infection,” “neonatal,” “mortality”) as well as microbiological and scientific research (e.g., “Klebsiella,” “plasmid,” “superbacteria”). Geographical and institutional entities such as “Brazil,” “WHO,” and “USP” also appeared, emphasizing G1’s role as an intermediary between scientific discourse and the general public.

Overall, while both corpora share core terminology (“antibiotic,” “bacteria,” “resistance”), their linguistic and semantic patterns diverge: Twitter emphasizes personal perceptions and misinformation, whereas G1 balances technical and popular narratives, sometimes adopting sensational angles to increase engagement. Thus, social media captures spontaneous discourse and emergent perceptions, whereas journalistic portals function as partial mediators of scientific communication.

Tuberculosis. This term was also analyzed across all three datasets (Twitter, G1, and Gov.br). On Twitter, seven rounds of clustering refinement were conducted, revealing clear, interpretable topics grouped into four main dimensions:

1. **Clinical and Preventive Aspects:** terms such as “tuberculosis,” “disease,” “diagnosis,” and “pneumonia,” reflecting educational campaigns and personal symptom narratives.

2. **Coinfections and Related Diseases:** including “AIDS,” “hepatitis,” “syphilis,” “leprosy,” and “influenza,” indicating users’ association of tuberculosis with other infectious diseases.

3. **Social and Sanitary Dimension:** words like “sanitation,” “rate,” and “health,” showing awareness of social determinants of the disease.

4. **Technical and Scientific Discussion:** presence of terms such as “Mycobacterium,” “bacillus,” and “tuberculosis,” associated with specialized or scientific discourse.

In G1, topics clustered into three main axes: (1) public health campaigns and prevention alerts emphasizing vaccination and early diagnosis; (2) scientific advances and epidemiological studies, including multidrug resistance; and (3) global and institutional perspectives, referencing organizations like WHO and the UN.

In Gov.br, topics mainly involved clinical guidelines, treatment protocols, and awareness campaigns focusing on free treatment and public access via Brazil’s Unified Health System (SUS).

Cross-Source Comparison. Table 3 summarizes the comparative findings between the two main

topics (Bacterial Resistance and Tuberculosis) and their discursive patterns across sources.

Commonalities across all sources include frequent references to prevention, diagnosis, and treatment — reinforcing the dominance of biomedical discourse even in informal contexts. Differences reveal complementary nuances: Twitter prioritizes personal and emotional experience, G1 acts as a translator of scientific knowledge, and Gov.br serves as a normative authority.

Both "bacterial resistance" and "tuberculosis" show intersection through antimicrobial resistance and HIV coinfection themes, recurrent across all datasets. These overlapping topics may serve as focal points for future automated surveillance studies integrating social and institutional data.

In summary, topic modeling successfully integrates heterogeneous textual data from social, journalistic, and governmental sources, allowing interpretation of evolving public narratives surrounding infectious diseases.

5 Conclusion

This work explored methods for identifying relevant subtopics concerning infectious diseases and antimicrobial resistance across distinct discursive domains using topic modeling applied to large-scale textual data. The research sought to understand how the public, the press, and official institutions communicate and represent issues related to antibiotic use, bacterial resistance, and some related disease such as tuberculosis.

By applying and comparing unsupervised approaches, particularly BERTopic, this study organized and interpreted heterogeneous textual data from three main sources: Twitter, G1, and Gov.br. The results demonstrate the potential of topic modeling to structure and interpret diverse linguistic data, uncovering discourse patterns that reflect public perception, misinformation, scientific dissemination, and institutional regulation.

Even in noisy environments such as Twitter, coherent and meaningful topics could be extracted, distinguishing personal, humorous, and misinformative content from technical or educational discussions. In journalistic and governmental sources, stronger semantic and terminological consistency was observed, reflecting G1's mediating function and Gov.br's normative character.

Comparative analysis between the "Bacterial Resistance" and "Tuberculosis" themes also revealed conceptual intersections — particularly antimicrobial resistance and HIV coinfection — recurrent across all datasets. Such convergence suggests high-value discursive nodes that can inform future automated monitoring efforts.

In summary, this research demonstrates that integrating social, journalistic, and institutional textual sources through topic modeling provides a promising analytical framework for understanding the flow of information about bacterial resistance and tuberculosis. The approach also offers valuable insights for public communication and surveillance policies in health domains.

Future work focuses on integrating Large Language Models (LLMs) to address data sparsity, moving beyond classical short-text models. Key strategies include leveraging LLMs for generative data augmentation—transforming short texts into richer pseudo-documents prior to modeling—and exploring prompt-based topic inference to achieve superior semantic coherence and interpretability compared to traditional embedding approaches.

Future work could include incorporating human-in-the-loop evaluation, such as word intrusion tasks or expert labeling, to provide more nuanced and reliable assessments of model performance.

6 Acknowledgments

The authors acknowledge financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Brazilian Government. 2023. Gov.br open data portal. https://www.gov.br/.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Grupo Globo. 2023. G1 news portal. https://g1.globo.com/.

Antônio Pereira De Souza Júnior, Pablo Cecilio, Felipe Viegas, Washington Cunha, Elisa Tuler De Albergaria, and Leonardo Chaves Dutra Da Rocha. 2022. Evaluating topic modeling pre-processing pipelines for portuguese texts. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 191–201.

Caitlin Doogan Poet Laureate, Wray Buntine, and Henry Linger. 2023. A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56(12):14223–14255.

Theme / Source	Twitter	G1 / Gov.br
Bacterial Resistance	Informal debate; veterinary use; misinformation	Outbreak reports; microbiological research
Tuberculosis	Campaigns; coinfections; social context	Clinical protocols; international health alerts

Table 3: Comparison of discourse patterns across data sources.

Ramanan Laxminarayan, Adriano Duse, Chand Wattal, Anita KM Zaidi, Heiman FL Wertheim, Nithima Sumpradit, Erika Vlieghe, Gabriel Levy Hara, Ian M Gould, Herman Goossens, and 1 others. 2013. Antibiotic resistance—the need for global solutions. *The Lancet infectious diseases*, 13(12):1057–1098.

Leland McInnes, John Healy, Steve Astels, and 1 others. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, and 1 others. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The lancet*, 399(10325):629–655.

Jim O'Neill. 2016. Tackling drug-resistant infections globally: final report and recommendations. Technical report, Review on Antimicrobial Resistance, London. Commissioned by the UK Government.

Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445.

Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *24th international joint conference on artificial intelligence, IJCAI 2015*, pages 2270–2276. AAAI Press/International Joint Conferences on Artificial Intelligence.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nádia FF da Silva, Marília Costa R Silva, Fabíola SF Pereira, João Pedro M Tarrega, João Vitor P Beinotti, Márcio Fonseca, Francisco Edmundo de Andrade, and André CP de LF de Carvalho. 2021. Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Brazilian Conference on Intelligent Systems*, pages 104–120. Springer.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

United Nations General Assembly. 2015. Transforming our world: the 2030 agenda for sustainable development. Resolution adopted by the General Assembly on 25 September 2015 (A/RES/70/1).

World Health Organization. 2024. Who bacterial priority pathogens list, 2024: Bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance. Technical report, World Health Organization, Geneva.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.

X Corp. 2025. X (formerly twitter).

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013a. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013b. A biterm topic model for short texts. *Proceedings of the 22nd International Conference on World Wide Web*, pages 1445–1456.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.