MIPS - Mapping the Relationship between Research, Innovation, and Society through Topic Modeling

Diogo Nolasco[1], **Jonice Oliveira**[1]

[1]Programa de Pós graduação em Informática, PPGI
Universidade Federal do Rio de Janeiro (UFRJ)
Cidade Universitária – Rio de Janeiro – RJ – Brasil

diogo.sousa@ppgi.ufrj.br, jonice@ic.ufrj.br

Abstract. *The proliferation of scientific data on the internet, including social networks, blogs, patents and articles, demands integrated analysis to guide strategic decisions. This data covers the life cycle of innovation, from scientific publications, through patents, to public perception of innovations. Understanding these dimensions allows you to better allocate resources and efforts in strategic areas that are well received by society. Although initiatives like Altimetry evaluate the specific impact of articles or authors, there is a gap in analyzing the interrelationships between ideas, discoveries, products and opinions over time. In this context, topic modeling emerges as a powerful tool for extracting and understanding themes in textual data. This work proposes an integrated technique to identify and analyze topics and their relationships in the scientific, technological and social spheres. The solution allows you to cross topics of different dimensions, revealing how an innovation is born, grows, turns into a commercial product and is perceived by society. Unlike current solutions that focus on a single dimension, this method offers a multidimensional view. The experiments carried out show the growth of research topics over time and how they were discussed by society, in addition to relating rumors and misinformation. The results indicate significant connections between the dimensions studied, highlighting the relationship of topics with milestones in the evolution of innovation. In the future, experiments will be conducted involving all three dimensions (science, technology and society), accompanied by the creation of visualizations to facilitate technological prospecting.*

1. Introduction

The management of public health crises in the digital age faces a dual challenge: addressing the biological spread of diseases and managing the unprecedented flow of information associated with them. The concept of "Infodemic" highlighted by the World Health Organization (WHO) [World Health Organization 2020], describes the overabundance of information (some accurate and some not) that makes it difficult for people to find trustworthy sources and reliable guidance when they need it. In this scenario, Health Informatics plays a crucial role not only in managing clinical data but also in monitoring the flow of medical knowledge from the laboratory to society.

Ideally, scientific evidence produced in academia (e.g., papers in PubMed) should rapidly translate into technological solutions (patents) and effective public health communication. However, a significant disconnect often exists between the scientific consensus and public perception. During recent health emergencies, such as the Zika virus epidemic (2015-2016) and the COVID-19 pandemic, this gap facilitated the spread of rumors and disinformation, potentially compromising adherence to preventive measures and vaccination campaigns. Understanding the semantic and temporal dynamics between scientific production and social discourse is, therefore, a critical requirement for modern epidemiological surveillance.

1.1. Problem Characterization and Motivation

The central problem addressed by this research is the difficulty in integrating and analyzing heterogeneous textual data from different analytical dimensions of the health ecosystem. Currently, data regarding medical discoveries, health technologies, and public opinion reside in isolated silos (Scientific Repositories vs. Social Media). Existing tools typically focus on single-dimension analysis—such as bibliometrics for papers or sentiment analysis for tweets—failing to capture the interrelationships between these spheres.

There is a lack of automated artifacts capable of answering complex questions such as: Did the scientific discovery regarding Zika and microcephaly immediately permeate social discussions? or Are the trending health topics on social media backed by scientific evidence, or are they semantically distant rumors? The motivation for this work stems from the need to provide health authorities and researchers with a computational framework to bridge this gap, enabling the use of social media data as "human sensors" for digital disease detection and rumor identification.

1.2. Objectives and Contributions

The general objective (GO) is to develop an integrated technique for identifying and analyzing the relationships between topics in Science and Technology (S&T) databases and social topics over time, in order to support strategic analyses for the country's development. Alongside the general objective this work seeks to answer the following research questions about the artifact:

- RQ1: Is it possible to establish social, scientific, and technological topics through the artifact?
- RQ2: Is it possible to establish relationships between topics extracted from different dimensions, or even from the same dimension but originating from different databases?
- RQ3: Is the artifact useful for analyzing existing relationships over a period of time?
- RQ4: Is the artifact useful for analyzing scientific research relationships at the topic level?

Applied to the context of Computing Applied to Health, the specific contributions of this work include:

1. Cross-Dimensional Topic Extraction: A method to extract coherent health topics from diverse sources (scientific abstracts, patent descriptions, and social media posts) using unsupervised machine learning (LDA).
2. Infodemiological Surveillance: The proposal of the Symmetric Kullback-Leibler Divergence (DSKL) as a metric to quantify the semantic distance between scientific evidence and social discourse, enabling the automated detection of "information voids" and potential rumors.

3. Validation in Real Health Scenarios: The rigorous application of the artifact in two major public health crises—the Zika Virus epidemic and the COVID-19 pandemic—demonstrating its ability to map the lifecycle of health information and identify disinformation trends.

By correlating the ”Scientific” dimension (validated knowledge) with the ”Social” dimension (public perception), this work contributes to the field of Digital Health by providing tools to mitigate the impact of disinformation and improve health communication strategies.

2. Methodology

This research adopted the Design Science Research (DSR) approach [Dresch et al. 2015], structured in iterative cycles to develop and validate the MIPS technique. The methodology focuses on processing heterogeneous textual data to track relationships between analytical dimensions and monitor them according to data such as the ones provided by public health crises.

2.1. Data Collection and Processing

To analyze the health ecosystem, data was collected from distinct dimensions: the Scientific Dimension, using crawlers to retrieve biomedical abstracts from PubMed (e.g., Zika Virus and COVID-19 corpora); the same was made in patent bases such as WIPO for the Techonological Dimension; and the Social Dimension, utilizing APIs to harvest public posts from Twitter/X, acting as ”human sensors” for real-time disease surveillance. All data underwent rigorous pre-processing, including tokenization and stop-word removal, to homogenize technical terminology with informal social language.

2.2. Topic Modeling

The core extraction engine relies on Latent Dirichlet Allocation (LDA) [Blei et al. 2003]. Unlike supervised methods that require pre-labeled medical datasets (often unavailable in early outbreak stages), LDA infers latent semantic structures from unlabelled text. To ensure the reliability of the extracted health topics, a Stability Analysis algorithm was implemented to mathematically determine the optimal number of topics (K), minimizing human bias in parameter selection. An automated labeling module based on n-gram ranking was used to interpret the probabilistic distributions (e.g., labeling a topic distribution as ”Microcephaly/Congenital Defects”).

2.3. Inter-Topic Relationship

The main innovation for health informatics is the Inter-Topic Relationship Module. To quantify the alignment between scientific evidence and public perception, the Symmetric Kullback-Leibler Divergence (DSKL) was employed [Kullback and Leibler 1951]. Since topics in LDA are probability distributions, DSKL measures the information loss when one distribution (e.g., Scientific Topic P) is used to approximate another (e.g., Social Topic Q).

$$D_{\text{SKL}}(P \parallel Q) = D_{\text{KL}}(P \parallel Q) + D_{\text{KL}}(Q \parallel P) = \sum_i \left(p_i \log \frac{p_i}{q_i} + q_i \log \frac{q_i}{p_i} \right) \quad (1)$$

This metric serves as a proxy for validity. A low DSKL distance indicates that social discourse is aligned with scientific findings (effective health communication). A high DSKL distance, particularly when associated with high social volume, flags an ”information void” or a potential health rumor (e.g., conspiracy theories regarding vaccines or vectors) that lacks semantic grounding in the scientific corpus. This allows for the automated detection of misinformation trends without manual fact-checking of every post.

3. Results and Discussion

The evaluation was conducted in various real-world scenarios, here we present those corresponding to two major global health crises: the Zika Virus epidemic (2015-2016) and the COVID-19 pandemic (2020). The experiments aimed to validate the artifact's ability to map the lifecycle of medical information and detect infodemic trends.

3.1. Zika Virus: Correlating Science and Society

This experiment processed 85,601 tweets and 1,769 PubMed articles across two time periods ranging from 2015 to 2016. The application of LDA with stability analysis and labeling extracted coherent topics in both dimensions. In the Scientific Dimension, topics evolved from general ”outbreak identification” (Period 1) to specific ”neurological inhibition” and ”sexual transmission” (Period 2). In the Social Dimension, topics mirrored these phases but focused on ”prevention/repellents,” ”travel warnings for pregnant women,” and ”concerns regarding the Rio 2016 Olympics.” [Nolasco and Oliveira 2020]

Using the DSKL metric, a heat map of semantic distances was generated. The results quantified the ”Social Impact” of medical discoveries:

- High Alignment (Low DSKL): There was a strong semantic correlation between scientific topics regarding microcephaly and social topics discussing ”birth defects.” This indicates that the core medical consensus successfully permeated public discourse. Similarly, scientific findings on sexual transmission rapidly generated corresponding social clusters warning about contagion risks.
- Information Voids (High DSKL): Conversely, technical topics regarding viral molecular structure remained semantically distant from social discussions, which is expected. However, specific social topics regarding local logistics showed high distance from all scientific topics, isolating technical knowledge from political discourse.
- The Monsanto/Larvicide Rumor: One specific social topic (labeled by the system as ”doctors expose monsanto linked pesticide”) exhibited the highest DSKL distance from the scientific baseline. This corresponded to a widespread conspiracy theory claiming microcephaly was caused by a larvicide (Pyriproxyfen), not the virus. The artifact correctly identified this as an outlier (a high-impact social narrative lacking scientific backing).
- Genetically Modified Mosquitoes: Another isolated topic linked the outbreak to the release of GM mosquitoes. The metric successfully distinguished this rumor from valid discussions on vector control.

3.2. COVID-19: Political Discourse and Technological Prospecting

In the COVID-19 scenario, the artifact was applied to two distinct ends:

- Health Technology Prospecting: Applied to patent databases (INPI/Espacenet), the unsupervised model automatically categorized innovations into ”Diagnosis” (PCR kits), ”Therapy” (vaccines), ”Protection” (masks), and ”Sanitization.” Topic 1 (”detection kit/method”) for example showed a 94% intersection with expert-curated categories, proving the tool’s utility for health technology assessment [Nolasco and Oliveira 2025].
- Political Infodemics: Analyzing the correlation between Presidential Speeches and social reaction, the Overlap Coefficient revealed that topics containing misinformation in official speeches (e.g., advocating unproven drug treatments like hydroxychloroquine or minimizing the virus as a ”little flu”) generated immediate and highly correlated echoes in social media. Unlike the Zika case, where rumors emerged from the bottom up, this analysis traced the top-down amplification of health misinformation [Nolasco and Oliveira 2021].

The results demonstrate that the MIPS applied to health data could serve as an effective Digital Disease Detection tool. By continuously monitoring the semantic distance between the ”Scientific Truth” (Repositories) and ”Public Perception” (Social Media), health authorities can identify: (i) Communication Gaps, where valid guidelines fail to reach the public; and (ii) Rumor Outbreaks, allowing for targeted interventions before disinformation compromises public health campaigns.

4. Conclusion

This research addressed the critical challenge of managing the ”Infodemic” accompanying modern public health crises. By proposing the MIPS (Mapping the Relationship between Research, Innovation, and Society) technique, this work provided a computational framework to bridge the gap between scientific evidence and public perception. The artifact proved effective in processing heterogeneous data streams—from biomedical literature to social media noise—offering a holistic view of how health information evolves, disseminates, and occasionally distorts into rumors.

4.1. Contributions to Health Informatics

The validation of the artifact across the Zika virus and COVID-19 scenarios resulted in three primary contributions to the field of Computing Applied to Health:

1. Automated Infodemiological Surveillance: The proposal validates the concept of ”human sensors” for digital disease detection. By extracting latent topics from social media and correlating them with scientific repositories (PubMed), the technique allows health authorities to monitor what the population understands about a disease in real-time, identifying communication gaps where scientific guidelines fail to permeate public discourse.
2. Semantic Metrics for Misinformation Detection: A major methodological contribution is the application of the Symmetric Kullback-Leibler Divergence (DSKL) as a proxy for information validity. The experiments demonstrated that social topics with high semantic distance from the scientific corpus are strong candidates for rumors. This was empirically proven by isolating the ”Monsanto/Larvicide” conspiracy theory during the Zika outbreak, enabling a semi-automated fact-checking pipeline that prioritizes high-impact falsehoods.

3. Health Technology Prospecting: Beyond social analysis, the artifact proved robust for analyzing the technological response to health crises. In the analysis of patent databases, the unsupervised model correctly categorized COVID-19 innovations into expert-defined groups (Diagnosis, Therapy, Protection), facilitating strategic decision-making for resource allocation in health innovation.

4.2. Future Directions

In conclusion, the MIPS technique offers a bridge between ”hard” medical data and ”soft” social signals. For future work, the integration of Large Language Models (LLMs) and Biomedical Knowledge Graphs is planned to refine the semantic granularity of the topics, allowing for the inference of causal relationships (e.g., “Does the rumor about vaccine X cause a drop in vaccination rates in region Y?”). This evolution aims to provide public health managers with predictive dashboards to anticipate and mitigate the effects of disinformation on population health.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dresch, A., Lacerda, D. P., Jr, J. A. V. A., Dresch, A., Lacerda, D. P., and Antunes, J. A. V. (2015). *Design Science Research*, pages 67–102. Springer International Publishing.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Nolasco, D. and Oliveira, J. (2020). Mining social influence in science and vice-versa: A topic correlation approach. *International Journal of Information Management*, 51.

Nolasco, D. and Oliveira, J. (2021). Topical rumor detection based on social network topic models relationship. *sol.sbc.org.brD Nolasco, J OliveiraiSys-Brazilian Journal of Information Systems, 2021•sol.sbc.org.br*, 14:5–27.

Nolasco, D. and Oliveira, J. (2025). Mirabr – a system for patent analysis. *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 838–845.

World Health Organization (2020). Covid-19 advice - mythbusters — who western pacific. https://www.who.int/westernpacific/emergencies/covid-19/information/mythbusters. Accessed: 2026-02-23.