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

Diogo Nolasco (student)[1], **Jonice Oliveira (advisor)**[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 growing proliferation of data on the Web and in multimedia systems, including scientific articles, patents, and social media discussions, covers the entire innovation lifecycle. However, the integrated analysis of these data to guide strategic decisions remains a challenge, given its heterogeneous and fragmented nature. The thesis proposes an integrated technique to identify, analyze, and map the relationships between topics present in the scientific, technological, and social dimensions over time. Using topic modeling, the solution reveals how innovation is born in research, transforms into technology, and is perceived by society, offering a multidimensional view that current solutions do not provide. Experiments demonstrate the method's ability to track the evolution of research themes and their social discussion, including the identification of rumors and disinformation. The results indicate significant connections between the dimensions, highlighting the approach's utility for technological foresight and impact analysis.*

1. Introduction

Science, technology, and innovation (S,T&I) are globally recognized as indispensable drivers for socioeconomic progress. In the Brazilian context, academic production shows significant figures, positioning the country among the top 15 producers of scientific articles worldwide. However, there is a latent paradox: the difficulty of translating this volume of knowledge into competitive industrial patents or agile responses to the pressing needs of the population.

This imbalance is often exacerbated by the fragmentation of Information Systems (IS). Data on basic research is isolated in repositories like PubMed; data on technological innovation resides in industrial property bases like INPI; and social perception regarding these themes is disseminated chaotically on social networks such as Twitter/X. The absence of an integrated view prevents public and private managers from identifying where science fails to communicate with society or where the market ignores emerging technological trends.

1.1. Problem Definition and Motivation

The core problem addressed in this thesis is the lack of automated and unsupervised methods that allow the mapping of semantic and temporal relationships between different innovation actors. Traditionally, this mapping is performed manually by experts through systematic reviews and bibliometric analyses based solely on citations. Such methods,

besides being costly, are static and unable to process the massive volume of unstructured data generated daily.

The motivation of this work lies in the need to provide strategic intelligence for the innovation ecosystem, enabling the detection of: (i) **research gaps**, where social demands find no echo in science; (ii) **rumors and misinformation**, where social discourse dangerously diverges from scientific consensus; and (iii) **thematic evolution**, tracking how theoretical concepts become technological artifacts.

1.2. Objectives

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 domains, or even from the same domain 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?

1.3. Research Context in Information Systems

This research is strategically aligned with the Grand Challenges in Information Systems in Brazil 2016-2026 (GranDSI-BR), as defined by the Brazilian Computer Society (SBC) [Boscarioli et al. 2017]. The proposal directly addresses gaps identified in two specific challenges:

Challenge 2: Information Systems in the Open World: This challenge highlights the need for systems capable of operating in environments characterized by heterogeneity, massive scale, and decentralized control. This thesis tackles this by proposing an artifact designed to handle unstructured data streams from the "open world" (web and social media). By integrating these uncontrolled sources with structured scientific repositories, the MIPS technique addresses the GranDSI-BR's call for methods that manage the "volume, velocity, and variety" of data to extract value and context from the open web.

Challenge 4: Sociotechnical View of Information Systems: The GranDSI-BR emphasizes that Information Systems are not socially neutral; they interact with and shape human behavior. This thesis is deeply rooted in this sociotechnical perspective. By developing tools to map public perception and detect disinformation (Cycle 4 of the research), the work provides technological support for understanding complex social phenomena. It addresses the ethical and social dimensions of IS by offering mechanisms to monitor how scientific information is distorted as it travels through social networks, contributing to the development of systems that foster a more informed and resilient society.

In summary, this work proposes a bridge between technical data mining capabilities and sociotechnical interpretation, offering a robust artifact to support decision-making in an increasingly complex and interconnected information ecosystem.

2. Research Methodology

This work adopts the Design Science Research (DSR) methodology, an approach focused on the creation and evaluation of artifacts to solve relevant practical problems while contributing to the knowledge base. Following the guidelines proposed by [Dresch et al. 2015], the research was structured into iterative cycles. Each cycle aimed to refine the proposed solution—the MIPS technique (Mapping the Relationship between Research, Innovation, and Society)—through the development, testing, and evaluation of computational artifacts applied to heterogeneous data sources.

The research process was driven by the need to integrate unstructured data from distinct domains—Science, Technology, and Society (STS)—to identify latent topics and their interrelationships over time. The methodology is bipartite: it encompasses the research process (organized in cycles) and the construction of the artifact (the technical framework).

2.1. MIPS (Proposed Artifact)

The primary artifact resulting from this research is an integrated computational technique composed of four modular stages: Data Collection, Topic Modeling, Inter-Topic Relationship Analysis, and Temporal Analysis. This framework allows for the ingestion of heterogeneous text data and the output of semantic correlations and evolutionary graphs. A visual model of the proposal is presented in Figure 1.

2.1.1. Data Collection and Pre-processing

Given the multidimensional nature of the problem, the collection module was designed to handle diverse data sources: scientific repositories (e.g., PubMed, arXiv) for the Scientific dimension; patent databases (e.g., INPI, Espacenet, WIPO) for the Technological dimension; and social media platforms (e.g., Twitter/X) for the Social dimension. Data collection utilized web crawlers and APIs to retrieve documents, which were then stored in a unified structure. Pre-processing routines were applied to homogenize the text, including tokenization, normalization (case folding, accent removal), and stop-word removal. Crucially, the method is language-independent, allowing for the analysis of multi-lingual corpora (e.g., Portuguese and English), which is essential for local vs. global comparative analyses.

2.1.2. Topic Modeling and Automatic Labeling

The core of the semantic extraction relies on Latent Dirichlet Allocation (LDA), a generative probabilistic model [Blei et al. 2003]. LDA assumes that documents are mixtures of topics, and topics are distributions of words. To overcome common limitations of unsupervised learning, the MIPS technique implements two specific enhancements:

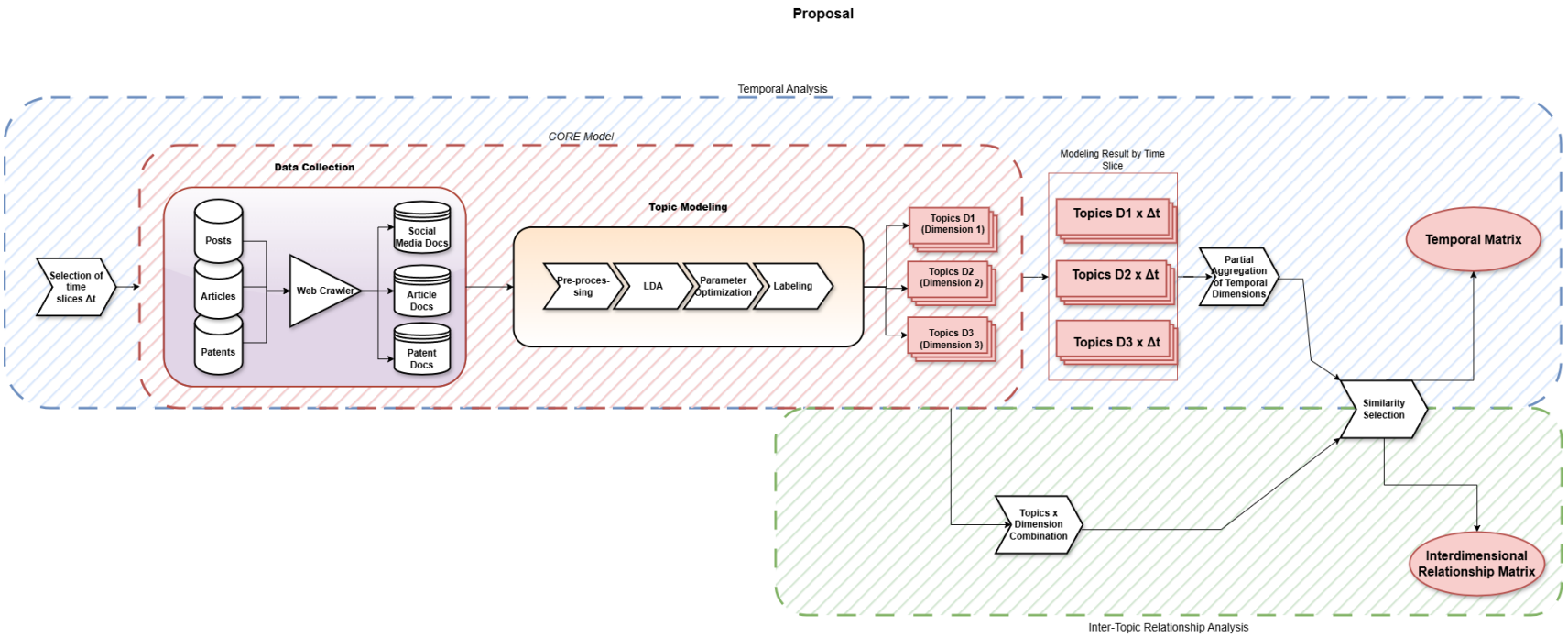

Figure 1. Schematic diagram of the Proposal

1. **Parameter Optimization via Stability Analysis:** To determine the optimal number of topics (K)—a critical parameter in LDA—the method employs a stability analysis algorithm. This involves generating random samples of the document collection and executing the LDA algorithm multiple times for a range of K values. The method selects the K that yields the most stable topic configurations across samples, ensuring that the extracted themes are robust features of the corpus rather than artifacts of a specific initialization.
2. **Automatic Labeling:** To make the probabilistic distributions interpretable by humans (and useful for decision-makers), a ranking algorithm is applied to assign semantic labels to each topic. The algorithm identifies candidate labels (n-grams) from the text and scores them based on their relevance to the topic's word distribution. This allows a topic defined by probabilities such as 0.04 gene, 0.02 dna, ... to be automatically labeled as "Genetics," facilitating the subsequent analysis of relationships between domains.

2.1.3. Inter-Topic Relationship Analysis

This module is responsible for identifying and quantifying the connections between different dimensions (e.g., "Is the scientific discussion on Zika Virus related to the social discussion on Microcephaly?"). Since topics in LDA are probability distributions, the method utilizes the Symmetric Kullback-Leibler Divergence (DSKL) [Kullback and Leibler 1951] to measure the semantic distance between topics from different domains. The distance

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

is calculated as the sum of the divergence of P from Q and Q from P, where P and Q are topics represented by probability distributions. A lower distance indicates a higher semantic similarity. This metric allows the construction of a Topic × Dimension matrix, revealing how a subject flows between Science, Technology, and Society. For scenarios involving non-probabilistic topic representations (e.g., manual tags or short speech transcripts), the method adapts by using an Overlap Coefficient [Vijaymeena and Kavitha 2016] to measure the intersection of terms, given by (where X and Y are term lists):

$$\text{overlapping}(X, Y) = \frac{|X \cap Y|}{\min\big(|X|, |Y|\big)} \quad (2)$$

2.1.4. Temporal Analysis

To map the trajectory of innovation and public perception, the methodology includes a temporal analysis module. The document collection is sliced into time windows (e.g., months or years). The topic modeling process is executed for each time slice, and an Evolutionary Graph is constructed. In this graph, nodes represent topics at a specific time t, and edges represent evolutionary transitions. An edge is established between a topic at time t and a topic at time t + 1 if their semantic similarity (calculated via document

intersection or probabilistic similarity) exceeds a defined threshold. This allows the identification of topic splits (divergence), mergers (convergence), and continuity, providing a visual representation of how a scientific idea transforms into a technological product or a social concern over time.

2.2. Research Cycles

Following the DSR guidelines, the development and evaluation of the MIPS technique were conducted in four iterative cycles, each addressing specific research questions and increasing in complexity.

Cycle 1: Science & Technology Identification: The first cycle focused on validating the topic modeling and temporal analysis algorithms within structured environments. Experiments were conducted using scientific conference data (SIGIR, SBBD) and patent databases. The objective was to verify if the artifact could automatically reconstruct the history of a research field and identify technological trends comparable to manual expert analysis.

Cycle 2: Social Topics and Subevent Detection: This cycle expanded the scope to the Social dimension, testing the artifact's ability to handle noisy, unstructured data from social media. The method was adapted to detect "subevents" within larger contexts, such as political protests ("Jornadas de Junho") and the early stages of the Zika epidemic. This validated the robustness of the data collection and pre-processing modules for non-scientific text.

Cycle 3: Impact Analysis (Cross-Domain): This critical cycle integrated the three dimensions. Using the Zika virus crisis as a case study, the research applied the DSKL metric to correlate scientific publications (PubMed) with social media discourse (Twitter). The goal was to measure the "Social Impact" of research—detecting if and when scientific findings (e.g., sexual transmission of the virus) permeated public discussion.

Cycle 4: Rumor Detection: The final cycle addressed a specific application of the relationship analysis: the identification of misinformation. By correlating "authoritative" topics (Scientific or Official Speeches) with "unverified" social topics, the method postulated that high semantic distance combined with high social volume could indicate rumors. This was validated against WHO rumor logs and fact-checking datasets regarding Zika and COVID-19.

This methodological structure ensured that the final artifact was not only theoretically sound but also empirically validated across different domains and real-world scenarios, directly addressing the challenge of managing knowledge in a networked society.

3. Related Work

To position the proposed artifact within the current state of the art, a systematic literature review was conducted following the guidelines of [Page et al. 2021]. The review focused on works published between 2019 and 2024 that utilized topic modeling for technological prospecting, scientific mapping, or social network analysis. The search was performed in the IEEE Xplore and SBC (Brazilian Computer Society) digital libraries, resulting in the analysis of 53 relevant studies after filtering.

The analysis of these works reveals a landscape where topic modeling is widely used but predominantly restricted to isolated domains (silos). The literature can be categorized into three main approaches: Single-Domain Analysis, Multi-Domain Analysis, and the identified Research Gap.

3.1. Single-Domain Analysis

The majority of existing research applies topic modeling to a single dimension of the innovation ecosystem, either Science, Technology, or Society, without crossing boundaries.

In the Social dimension, studies like [de S. Costa et al. 2022] utilize Latent Dirichlet Allocation (LDA) to analyze specific events, such as the International Women's Day on Twitter. Similarly, [de Sousa and Becker 2022] employed BERTopic to investigate vaccine hesitation regarding COVID-19 in Brazil and the USA. While these studies successfully identify social sentiment and temporal peaks of discussion, they analyze the "social bubble" in isolation. They lack a mechanism to automatically validate public opinion against scientific consensus or to measure the lag between a scientific discovery and its social repercussions.

In the Scientific and Technological dimensions, works such as [Parsons and Khuri 2020] focus on extracting trends from computer science education papers, while [Maskittou et al. 2022] apply Singular Value Decomposition (SVD) specifically to patent databases. These approaches are effective for internal mapping (e.g., identifying "hot topics" within a conference) but fail to capture the external impact of these innovations. For instance, they cannot answer if a surge in patents translates to increased societal awareness.

3.2. Multi-Domain Analysis

Recent efforts have attempted to bridge these domains, though often with limitations regarding automation or temporal granularity.

[Li et al. 2022] proposed a framework combining patent analysis with web news mining to track the development of Perovskite solar cells. Their work represents a significant step towards multidimensional analysis, comparing technological evolution with public expectations. However, the correlation between the two dimensions was performed through qualitative inference rather than an automated metric, limiting scalability to other domains.

[Machado et al. 2023] utilized Generative AI and topic modeling to analyze the convergence between Science (articles) and Technology (patents) in the context of virtual worlds. While successful in identifying seven convergence topics, the study did not incorporate the social dimension (public perception) and did not provide a mechanism for tracking the temporal evolution of these relationships.

3.3. Synthesis and Research Gap

The critical analysis of the literature reveals a significant gap: the lack of an integrated technique capable of (1) handling heterogeneous data (Science, Technology, and Society) simultaneously, (2) performing automated correlation between these domains, and (3) tracking the temporal evolution of these relationships.

Most existing tools operate on the assumption of independent topics or rely on manual curation to link different datasets. Furthermore, few works address the issue of disinformation (rumors) by contrasting social discourse with scientific databases at the topic level.

Table 1 summarizes the comparison between the proposed MIPS technique and key related works, highlighting the unique combination of features offered by this thesis.

Table 1. Comparison of the proposed approach (MIPS) with related works

Citation	Algorithm	Data Sources	Multi-dim. Data?	Temporal Analysis?	Automated Rel. Analysis?
MIPS (Proposal)	**LDA + DSKL**	**Articles, Patents, Social**	**Yes**	**Yes**	**Yes**
Liu et al. (2021)	DTM	IEEE Articles	No	Yes	No
Li et al. (2022)	hLDA	Patents, News	Yes	No	No (Manual)
Machado et al. (2023)	GenAI	Patents, Articles	Yes	No	No
Sousa; Becker (2022)	BERTopic	Twitter	No	Yes	No
Zhang et al. (2023)	LDA	Articles	No	No	No

As shown in Table 1, while methods like Dynamic Topic Models (DTM) used by [Liu et al. 2021] handle temporal evolution effectively, they are restricted to a single data type. Conversely, works that mix data types often lack the temporal or automated relationship components. The MIPS technique fills this gap by providing a unified framework where topics are not only extracted but mathematically correlated across dimensions and time, enabling a holistic view of the innovation lifecycle.

4. Results and Discussion

MIPS evaluation followed the four research cycles , verifying its effectiveness in extracting and labeling topics from heterogeneous sources , tracking temporal evolution, and uncovering cross-domain relationships and misinformation.

4.1. Validation of Topic Extraction and Temporal Analysis (Cycle 1)

The first phase of evaluation focused on the robustness of the topic modeling and temporal analysis modules using ”ground truth” data from scientific conferences and patent databases.

4.1.1. Scientific Evolution: The SIGIR and SBBD Case Studies

MIPS reconstructed the history of SIGIR (Special Interest Group on Information Retrieval) and SBBD (Simpósio Brasileiro de Banco de Dados) conferences to validate temporal analysis. The objective was to reconstruct the history of these research fields automatically and compare the results with manual longitudinal studies conducted by domain experts [Smeaton et al. 2003, Kauer 2013]. Figure 2 illustrates the resulting Temporal Evolutionary Graphs.

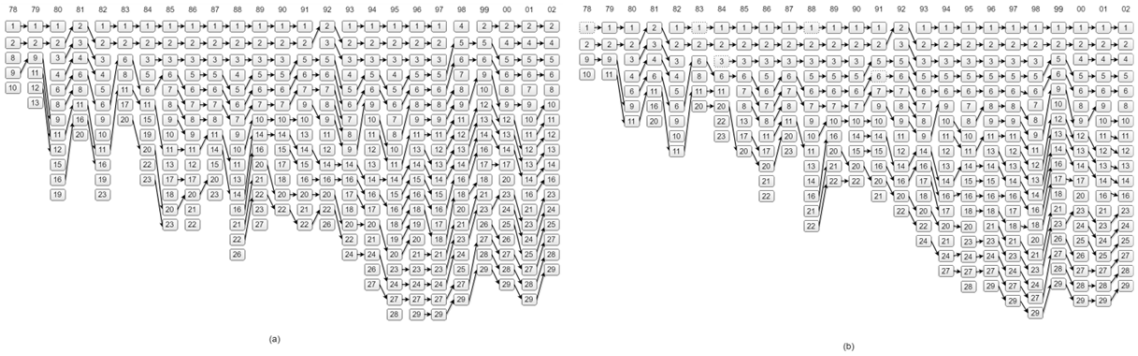

Figure 2. Comparison of Temporal Evolutionary Graphs between (a) ground truth and (b) MIPS

For the SIGIR dataset (1978–2002), MIPS identified 232 evolutionary transitions, achieving 73% coverage of the manual study by [Smeaton et al. 2003] . Crucially, the automated method revealed granular connections missed by experts; while human analysis treated ”Database Systems” as a monolithic block, MIPS detected specific branching into sub-topics like ”Query Optimization” and ”Object-Oriented Databases”.

Similarly, for SBBD, the system detected 76 transitions. Comparison with Kauer’s study [Kauer 2013] demonstrated the sensitivity of the artifact in identifying the late-1990s emergence of ”Data Mining” as a field distinct from traditional database storage. These results confirm that the Evolutionary Graph can effectively map the ”genealogy” of research themes without human intervention.

4.1.2. Technological Prospecting: COVID-19 Patents

To evaluate the technological dimension, COVID-19 patents filed between January and August 2020 were analyzed to replicate expert categorizations [Falciola and Barbieri 2022]: Diagnosis, Therapy, Protection, and Sanitization . MIPS identified eight topics (Table 2) mapped to these categories with high precision via automatic labeling and document intersection. For instance, Topics 1 (Diagnosis), 4 (Sanitization), and 6 (Protection) yielded intersections of 94%, 98%, and 91%, respectively . Additionally, the system identified nuanced sub-topics within ”Therapy,” distinguishing ”Traditional Chinese Medicine” (Topic 2) from ”Vaccine development” (Topic 3). This granularity is essential for strategic decision-making, enabling stakeholders to pinpoint specific technological niches rather than broadly active fields

Table 2. Patent Topics with corresponding Overlaps

Topic	Labels	Manual Topics Overlap (%)
1	“detection kit”, “coronavirus detecting”	Diagnosis (94%), Therapy (5%)
2	“chinese traditional medicine”, “pneumonia”	Therapy (88%), Diagnosis (10%)
3	“sarscov2”, “fast vaccine development”	Therapy (73%), Diagnosis (16%)
4	“disinfection”, “air disinfectant”	Sanitization (98%), Protection (1%)
5	“preventing”, “treating”, “inflammation”	Therapy (76%), Protection (20%)
6	“protective mask”, “covid face mask”	Protection (91%), Sanitization (5%)
7	“hand sanitizer”, “sanitizer dispenser”	Sanitization (77%), Protection (9%)
8	“antivirus”, “antiviral”, “immunity”	Therapy (80%), Protection (9%)

4.2. Identification of Social Sub-events (Cycle 2)

The second cycle evaluated the system's ability to process noisy, informal text from social media (Twitter/X) to identify sub-events within broader crises. This was tested in two different settings: the *Jornadas de Junho* protests in Brazil and the onset of the Zika virus epidemic. A quantitative metric, Relevance ($\beta > \alpha$), was adopted, where β represents topics corroborated by official news media and α represents detected topics with no corresponding news evidence (potential noise).

In the *Jornadas de Junho* case, the system analyzed 432,975 tweets collected between June and August 2013 across two major protest peaks (16–18/06/2013 and 19–21/06/2013). MIPS identified 14 sub-events in the first period and 20 in the second, with low noise levels ($\alpha = 1, \beta = 13$ and $\alpha = 2, \beta = 18$, respectively), showing that the artifact could recover relevant protest-related sub-events from noisy social media data.

In the Zika epidemic scenario, the system processed over 85,000 tweets and successfully identified sub-events such as the WHO's declaration of emergency (Topic 3), travel warnings for pregnant women (Topic 4), and the cancellation of Olympic preparations due to the virus (Topic 2). Taken together, the two scenarios yielded $\beta \gg \alpha$, with 31 relevant sub-events against only 3 noise topics, demonstrating that MIPS effectively filters social noise while preserving societally meaningful events.

4.3. Cross-Domain Impact Analysis: Science vs. Society (Cycle 3)

The most significant contribution of the MIPS technique is the ability to correlate topics across dimensions. This experiment utilized the Zika virus crisis (2015-2016) to analyze the relationship between Scientific Research (PubMed articles) and Public Opinion (Twitter/X).

Using the Symmetric Kullback-Leibler Divergence (DSKL), a heatmap of semantic distances was generated (as can be seen in Table 3). The analysis was divided into two periods: Period 1 (Outbreak Start) and Period 2 (Peak/Decline).

The results revealed strong semantic correlations (low DSKL distance, represented in green in the heatmaps) between specific scientific breakthroughs and social discussions.

Microcephaly Link: Scientific Topic 1 in Period 2 ("Congenital fetal malformations") showed a low semantic distance to Social Topic 1 ("birth defect microcephaly"). This indicates that the scientific consensus on the link between Zika and birth defects successfully permeated the public discourse.

Transmission Vectors: Scientific research on alternative transmission methods (saliva, sexual transmission) correlated significantly with social topics warning about these risks.

Conversely, high semantic distances (red zones) highlighted gaps. Some scientific topics regarding specific molecular interactions of the virus had no corresponding social topic, which is expected. However, some social topics, particularly those regarding local logistics or political complaints, showed high distance from all scientific topics, isolating technical knowledge from political discourse.

The temporal aspect of this analysis showed that while scientific topics became

Table 3. Alignment between Social and Scientific Topics in different Periods

Period 1 – May/15 to Feb/16						
		Scientific Topics				
		1	2	3	4	5
Social Topics	1	4.02	2.28	4.71	1.03	2.48
	2	2.44	1.99	1.33	1.47	1.98
	3	1.21	1.34	3.17	4.93	1.61
	4	2.86	3.96	4.41	2.87	1.09
	5	4.58	3.98	2.71	4.13	2.75
	6	1.46	4.30	3.14	1.47	2.16
	7	0.74	2.34	0.67	2.71	1.72
	8	3.44	4.26	4.92	3.74	2.80

Period 2 – Mar/16 to Dec/16								
		Scientific Topics						
		1	2	3	4	5	6	7
Social Topics	1	4.90	4.46	2.68	0.52	2.11	4.71	1.38
	2	4.61	1.12	3.55	0.70	3.75	4.22	2.14
	3	2.44	4.60	4.87	0.92	3.01	0.54	2.49
	4	2.60	2.18	2.65	2.50	2.97	2.89	2.61
	5	0.51	3.43	1.37	3.98	4.49	1.30	2.65
	6	3.13	4.50	3.96	1.93	3.92	1.41	3.27
	7	1.18	1.11	1.11	4.25	3.03	3.43	3.59
	9	3.54	3.59	4.70	2.45	2.84	3.84	4.01

more specialized over time (moving from general ”outbreak” to ”neurologic inhibition”), social topics remained focused on prevention and immediate consequences, reacting to scientific findings with a lag.

4.4. Application in Disinformation and Rumor Detection (Cycle 4)

The final evaluation explored the hypothesis that high semantic distance between a social topic and the ”authoritative” domain (Science) could indicate the presence of rumors or disinformation.

4.4.1. The Zika Rumors Case

By analyzing the topics with the highest DSKL distance from the scientific corpus (i.e., the ”reddest” topics in the previous analysis), the system isolated clusters of misinformation. In this case the results were compared with a ground truth of rumor topics from the World Health Organization [World Health Organization 2016]. Table 4 shows the correspondence between MIPS rumor detected topics and rumor topics as classified by WHO.

Monsanto Rumor: One specific social topic (labeled ”doctors expose monsanto linked pesticide”) had an extremely high semantic distance from all PubMed topics. This corresponded to a widespread conspiracy theory claiming that microcephaly was caused by a larvicide, not the virus. The MIPS technique successfully flagged this as an outlier—a topic with high social volume but zero scientific grounding.

Genetically Modified Mosquitoes: Another isolated topic linked the outbreak to the release of GM mosquitoes.

The system’s ability to mathematically isolate these topics suggests the utility of topic correlation as a topic analysis tool by fact-checking for example. If a high-volume social topic has no semantic overlap with the scientific corpus, it warrants immediate investigation by health authorities.

4.4.2. Political Discourse and COVID-19

A second case study analyzed the correlation between Presidential Speeches (Official Discourse) and Social Media reaction during the early months of the COVID-19 pandemic in Brazil (2020). Here, the ”authoritative” source was the transcription of speeches, and the metric used was the Overlap Coefficient, due to the smaller vocabulary of the speeches.

Table 4. Comparison: Rumor Topics (Period, Topic) vs. WHO Classified Rumors

Rumor Topics	WHO Classified Rumors
Monsanto pesticides & microcephaly (P1, T1)	Pyriproxyfen causes microcephaly
Tata "Zica" car naming confusion (P1, T7)	Vaccines cause microcephaly
First cases in several countries (P2, T5)	Zika symptoms are same as seasonal flu
Mosquito mutation causing spread (P2, T6)	Bacteria in male mosquitoes spread Zika
Repellent info and propaganda (P2, T6)	Certain repellents work better
–	GM mosquitoes linked to Zika in Brazil
–	Sterilized males contribute to spread

The analysis involved expert annotation of speech topics as containing "misinformation" (e.g., advocating unproven drug treatments like hydroxychloroquine) or "correct information." The results showed a high correlation between the President's speech topics labeled as misinformation and subsequent social media topics.

The "Gripezinha" Effect: When the speech minimized the virus (comparing it to a "little flu"), a corresponding social topic emerged immediately with high overlap.

Hydroxychloroquine: Topics advocating for unproven drugs in the speeches generated strong echoes in social media topics (defending the usage), which, if cross-referenced with the scientific domain (as in the Zika experiment), would show high distance from the medical consensus.

The comparison between the Zika case (Science vs. Society) and the COVID case (Politics vs. Society) highlights the versatility of the MIPS technique. In the Zika case, the lack of correlation with science flagged rumors. In the COVID case, the strong correlation with political speeches traced the source and amplification of information (or disinformation).

4.5. Summary of Findings

The experimental results across the four cycles provide empirical evidence to answer the Research Questions (RQ) defined at the inception of this work. The MIPS artifact demonstrated robustness and versatility in the following aspects:

Regarding RQ1: Is it possible to establish social, scientific, and technological topics through the artifact? The experiments confirmed that the proposed pipeline—integrating crawler-based data collection, LDA with stability analysis, and automatic labeling—successfully extracted coherent topics across all tested dimensions [Nolasco and Oliveira 2025, Nolasco and Oliveira 2018a].

Regarding RQ2: Is it possible to establish relationships between topics extracted from different areas or even from the same area but from different bases? Yes. The introduction of the Symmetric Kullback-Leibler Divergence (DSKL) and Overlap Coefficient metrics proved effective in quantifying semantic relationships between heterogeneous sources [Nolasco and Oliveira 2018b, Nolasco and Oliveira 2020a].

Regarding RQ3: Is the artifact useful for analyzing relationships existing over a period of time? The Temporal Analysis module was validated through the construction of Evolutionary Graphs [Nolasco and Oliveira 2025, Nolasco and Oliveira 2019].

Regarding RQ4: Is the artifact useful for analyzing scientific research relation-

ships at the topic level? The artifact provided a granular view of scientific relationships that goes beyond simple keyword matching. By modeling research as probability distributions over vocabulary, the system identified subtle thematic connections (e.g., the link between Dengue and Zika research vectors) and validated them against authoritative agendas, such as the WHO R&D Blueprint. The ability to isolate "outlier" topics (those with no scientific relationship) also proved useful for the inverse analysis: detecting non-scientific rumors [Nolasco and Oliveira 2020b, Nolasco and Oliveira 2021].

5. Scientific Production and Academic Products

The development of this doctoral research extended beyond the theoretical proposal of the MIPS technique, resulting in a consistent production of scientific knowledge, technological artifacts, and human resources formation. The validation cycles described in this work were subjected to rigorous peer review, achieving recognition in high-impact journals and renowned international conferences.

5.1. Publications in Journals and Conferences

The core contributions of this thesis were published in top-tier venues. Notably, it resulted in two best paper awards in two conferences (IEEE/DASC and BraSNAM/CSBC) and the research resulted in two articles in JCR Q1 journals, demonstrating the global relevance of the findings. Table 5 summarizes the main bibliographic production associated with the thesis.

Table 5. Summary of Scientific Production (Indicators represent JCR/Qualis)

Type	Publication Title	Venue / Journal	Indicators
Journal	Mining social influence in science and vice-versa: A topic correlation approach	Int. J. Inf. Manage. (IJIM)	Q1 / A1
Journal	Subevents detection through topic modeling in social media posts	Future Gen. Comp. Syst. (FGCS)	Q1 / A1
Journal	Topical Rumor Detection based on Social Network Topic Models Relationship	iSys - Brazil. J. Inf. Syst.	A4
Conf.	MiraBR – A System for Patent Analysis	SBSI 2025	A4
Conf.	Intelligent subevent detection based on social network data	IEEE DASC	B4
Conf.	Publish or Post: Identification of Influences between Science and Society	VLDB Workshop	A1
Conf.	A Study of Rumor Detection based on Social Network Topic Models Relationship	BraSNAM (CSBC)	B1

5.2. Technological Products and Open Source Artifacts

In alignment with the principles of Open Science and reproducibility, the computational artifacts developed in this thesis were made available to the community.

MiraBR System: A dedicated system for patent analysis and technological prospecting, presented at SBSI 2025 [Nolasco and Oliveira 2025].

MIPS: The source code for the "MIPS" and the implementations of the DSKL metric and LDA parameter optimization are available on GitHub, allowing other researchers to replicate the experiments or apply the MIPS technique to new domains.

Implemented components: scraping/ingestion scripts (e.g., PubMed for articles; INPI/EPO/WIPO for patents; Twitter for social data), utilities for text cleaning/normalization, labeling heuristics.

COVID-19 Knowledge Graph (system/dataset): A graph visualization system to explore the knowledge graph built using the methods and data of the thesis. It is documented in [Cerri and Nolasco 2023].

5.3. Human Resources Formation and Mentorship

The methodology developed in this thesis served as the foundation for the training of new researchers, demonstrating its transferability and educational value.

Scientific Initiation Co-supervision: The author co-supervised the undergraduate research project of student Andre Cerri (UFRJ), titled "Integração de dados científicos sobre Covid-19 através de Grafos de Conhecimento" [Cerri et al. 2022]. This mentorship applied the topic modeling concepts developed in Cycle 1 and Cycle 3 of this thesis to build a semantic graph of the pandemic, resulting in two best work award in the UFRJ Academic Integration Week.

The author delivered invited lectures at prestigious institutions such as INPI, Fiocruz, UFRJ, and UFRRJ, and served as a judge for the NCE/UFRJ Hackathon. Furthermore, this doctoral research was integrated into major international and national initiatives, enabling the validation of the MIPS framework within multidisciplinary teams:

ZIKAlliance: A global consortium where the social sensing techniques (Cycle 2 and 3) were applied to understand the social impact of the Zika Virus in Brazil.

Latin America Covid-19 Social Sciences Initiative: Application of the MIPS technique to extract scientific data from non-indexed sources during the pandemic.

Extracting Knowledge from Big Social Data: A project focused on managing high-volume social data during sanitary crises, which provided the environment for testing the scalability of the proposed algorithms.

In conclusion, this thesis not only achieved its specific objectives but also established a validated framework for the analysis of the Science-Technology-Society triad, with proven impact across academic, technological, and social dimensions.

6. Conclusion

This work addressed the challenge of analyzing multidimensional relationships between Science, Technology, and Society (STS) by proposing MIPS, an integrated topic-modeling technique that bridges heterogeneous sources such as scientific articles, patents, and social media posts. The resulting framework makes it possible to map how knowledge evolves, becomes innovation, and reaches public perception over time.

The development and evaluation of MIPS answered the research questions defined at the outset. The pipeline, combining parameter stability analysis and automatic labeling, showed that coherent topics can be extracted across distinct domains (RQ1). The use of Symmetric Kullback-Leibler Divergence (DSKL) and the Overlap Coefficient enabled the quantification of semantic relationships between heterogeneous sources (RQ2), revealing, for example, the lag between scientific discoveries on Zika and their discussion on social

media. In addition, the Temporal Analysis module supported the visualization of topic evolution across time (RQ3), reconstructing research trajectories in the SIGIR and SBBD case studies and identifying technological trends in patent data. Finally, topic-level analysis helped isolate specific themes and distinguish scientifically grounded discussions from misinformation (RQ4).

From an Information Systems perspective, this thesis contributes to the relationship between people, processes/organizations, and technologies: it supports the analysis of social perception and misinformation, improves strategic monitoring and coordination across research and innovation environments, and provides a computational artifact for integrating heterogeneous textual sources. The results, however, should be interpreted in light of some limitations, including dependence on the selected datasets and case studies, the challenges of integrating highly heterogeneous sources, and the sensitivity of topic modeling to data quality, domain specificity, and multilingual variation.

6.1. Future Work

Future work includes incorporating Knowledge Graphs, Ontologies, and Large Language Models to improve topic labeling, semantic disambiguation, and explanation of cross-domain relationships, as well as expanding the approach to multilingual and international datasets.

References

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

Boscarioli, C., Araujo, R. M., and Maciel, R. S. P., editors (2017). *I GranDSI-BR – Grand Research Challenges in Information Systems in Brazil 2016–2026*. Brazilian Computer Society (SBC), Special Committee on Information Systems (CE-SI). ISBN 978-85-7669-384-0.

Cerri, A. and Nolasco, D. (2023). *KNOWLEDGE GRAPH FOR COVID-19: A FUSION APPROACH - SIAC UFRJ*, pages 385–385.

Cerri, A., Nolasco, D., and Oliveira, J. (2022). *Integração de dados científicos sobre Covid-19 através de Grafos de Conhecimento - SIAC UFRJ*, pages 401–401.

de S. Costa, G., Couto, D. C. C., Junior, A. F. L. J., and Lobato, F. M. F. (2022). Feminismo e redes sociais online: uma análise de tweets sobre o dia internacional da mulher. *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 169–180.

de Sousa, A. M. and Becker, K. (2022). Comparando os posicionamentos a favor/contra a vacinação covid nos estados unidos da américa e no brasil. *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 65–77.

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.

Falciola, L. and Barbieri, M. (2022). Searching and analyzing patent-relevant covid-19 information. *World Patent Information*, 68:102094.

Kauer, V. A. (2013). Evolução dos temas de interesse do sbbd ao longo dos anos. *Simpósio Brasileiro de Banco de Dados*, pages 1–6.

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

Li, X., Xie, Q., and Huang, L. (2022). Identifying the development trends of emerging technologies using patent analysis and web news data mining: The case of perovskite solar cell technology. *IEEE Transactions on Engineering Management*, 69:2603–2618.

Liu, Y., Wang, J., Qian, Y., Jiang, Y., Sun, J., and Chai, Y. (2021). Dynamic topic model for tracking topic evolution and measuring popularity of scientific literature. In *2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC)*, pages 315–320.

Machado, L. B., Moresi, E. A. D., Ferneda, E., and Prado, H. A. D. (2023). Bib-ana: a tool to detect emerging technologies through topic modeling and generative artificial intelligence. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6.

Maskittou, M., Haddadi, A. E. L., and Routaib, H. (2022). Intelligent technology management based on patent topic modeling. In *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS)*, pages 1–4.

Nolasco, D. and Oliveira, J. (2018a). Intelligent subevent detection based on social network data. In *Proceedings - 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing*, volume 2018-January.

Nolasco, D. and Oliveira, J. (2018b). Publish or post: Identification of influences between science and society through intelligent systems. *ceur-ws.orgD Nolasco, J OliveiraBiDu-Posters@ VLDB, 2018•ceur-ws.org*.

Nolasco, D. and Oliveira, J. (2019). Subevents detection through topic modeling in social media posts. *Future Generation Computer Systems*, 93:290–303.

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

Nolasco, D. and Oliveira, J. (2020b). A study of rumor detection based on social network topic models relationship. In *sol.sbc.org.br*, pages 166–177.

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.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Parsons, S. and Khuri, N. (2020). Discovery of research trends in computer science education on ethics using topic modeling. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 885–891.

Smeaton, A. F., Keogh, G., Gurrin, C., McDonald, K., and Sodring, T. (2003). Analysis of papers from twenty-five years of sigir conferences: what have we been doing for the last quarter of a century? In *SIGIR Forum*, volume 37, pages 49–53.

Vijaymeena, M. and Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

World Health Organization (2016). Who – dispelling rumours around Zika and complications. `http://www.who.int/emergencies/zika-virus/articles/rumours/en/`. Accessed: 2026-02-23.