Demographic and Geographic Drivers of Scientific Trends: Covariate Effects in Canadian NSERC Proposals

Shirin Tavakoli Kafiabad

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

October 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Shirin Tavakoli Kafiabad**

Entitled: **Demographic and Geographic Drivers of Scientific Trends: Covariate Effects in Canadian NSERC Proposals**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Name of the Chair

_____ External Examiner
Dr. Name of External Examiner

_____ Examiner
Dr. Name of Examiner One

_____ Supervisor
Dr. A. Schiffauerova

Approved by _____
Dr. C. Wang, Chair
Department of Concordia Institute for Information Systems Engineering

_____ 2025

_____
Dr. M. Debbabi , Dean
Faculty of Engineering and Computer Science

Abstract

Demographic and Geographic Drivers of Scientific Trends: Covariate Effects in Canadian NSERC Proposals

Shirin Tavakoli Kafiabad

Optimizing national scientific investment requires a clear understanding of evolving research trends and the demographic and geographical forces shaping them, particularly in light of commitments to equity, diversity, and inclusion. This thesis investigates how researcher gender and provincial location influence the prevalence and evolution of research topics over 18 years (2005–2022) of proposals funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). To address this objective, we conducted a comprehensive comparative evaluation of three topic modeling approaches: Latent Dirichlet Allocation (LDA), Structural Topic Modeling (STM), and transformer-based BERTopic. A key innovation is the COFFEE pipeline, a novel Python tool that enables robust covariate effect estimation for BERTopic. This advancement addresses a significant gap, as BERTopic lacks a native function for covariate analysis, unlike the probabilistic STM. Our findings highlight that while all models effectively delineate core scientific domains, BERTopic outperformed by consistently identifying more granular, coherent, and emergent themes, such as the rapid expansion of artificial intelligence. Additionally, the covariate analysis, powered by COFFEE, confirmed distinct provincial research specializations and revealed consistent gender-based thematic patterns across various scientific disciplines. These insights offer a robust empirical foundation for funding organizations to formulate more equitable and impactful funding strategies, thereby enhancing the effectiveness of the scientific ecosystem.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Andrea Schiffauerova and Dr. Ashkan Ebadi, for their invaluable advice, continuous support, consistent encouragement, and patience during my master's study.

A special thanks to my fellow lab mates and great friends at Concordia University, particularly, Sadaf Khademi, Shahin Heidarian, Parastoo H.Meybodi, Farnoush Bayatmakou.

I would also like to extend my heartfelt thanks to my family, especially my mother, father, and my brother, Shayan, for their unwavering love, support, and encouragement. They have been a constant source of inspiration and motivation throughout my journey, and I am incredibly grateful for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's rapidly evolving global landscape, understanding the forces that drive scientific inquiry and the dynamics of research funding is more crucial than ever. As global investment in science and technology continues to increase Ebadi, Tremblay, Goutte, and Schiffauerova (2020), there is a growing imperative for clear data-driven insights into research trends and funding dynamics, particularly to address equity, diversity, and inclusion. This thesis aims to uncover how researchers' gender and provincial location influence the prevalence and evolution of funded research topics over 18 years (2005-2022) of proposals from the Natural Sciences and Engineering Research Council of Canada (NSERC). These insights are essential to ensure that substantial investments translate into meaningful social and economic returns, contributing to equitable resource allocation Stephan (2015).

To effectively navigate and extract meaningful insights from this extensive textual funding dataset, this thesis employs advanced computational techniques in Natural Language Processing (NLP) and Artificial Intelligence (AI), with a specific focus on topic modeling (TM). Topic modeling enables systematic discovery of latent themes within large corpora and tracking of how these themes evolve across demographic and temporal dimensions Blei and Lafferty (2006); Rosen-Zvi, Griffiths, Steyvers, and Smyth (2012). This approach offers a powerful lens for examining the evolving landscape of scientific research. Methodologically, first, we undertake a rigorous comparative analysis of three prominent topic models: 1) Latent Dirichlet Allocation (LDA) Blei, Ng, and Jordan (2003), which utilizes probabilistic inference to identify patterns in the data, 2) Structural Topic

Modelling (STM) Roberts, Stewart, Tingley, and Airoldi (2013), which builds on LDA by incorporating document-level covariates, allowing for a more nuanced understanding of thematic variations influenced by external factors, and 3) an advanced topic modelling approach based on Bidirectional Encoder Representations from Transformers (BERT), which leverages transformer-based contextual embeddings to capture semantic nuances Devlin, Chang, Lee, and Toutanova (2019); Grootendorst (2020). This multimodel comparison is crucial as it enables us to assess how different underlying modeling assumptions, ranging from probabilistic inference in LDA and STM to semantic space analysis through embeddings and clustering in the BERT-based approach, affect the coherence of identified topics and the interpretability of subsequent downstream analyses.

While identifying which research initiatives are receiving funding is crucial, gaining a comprehensive understanding of who conducts this research and where it takes place provides a more holistic view Bornmann, Mutz, and Daniel (2007a); Ogden (2019); Witteman, Hendricks, Straus, and Tannenbaum (2019). The existing literature highlights persistent gender and geographic disparities in scientific funding and participation Asheim, Grillitsch, and Trippl (2016); Breschi, Lenzi, Lissoni, and Vezzulli (2010); Rodríguez-Pose (2018). However, there is a critical gap in systematically quantifying how these researcher characteristics influence the prevalence of specific research topics within large-scale funding datasets and how these relationships might change over time, potentially reflecting systemic inequities or shifts in research priorities. This study addresses this gap by leveraging advanced topic modeling and a novel covariate effect estimation pipeline.

Addressing this gap, a central and novel component of this study involves a comparative estimate of the quantitative influence of researcher characteristics, particularly gender and geographic locations, on the prevalence of different research topics in both the STM and the BERTopic. This part of the research leverages two prominent topic modeling approaches: Structural Topic Modeling (STM) Roberts et al. (2013) as a baseline, which integrates metadata directly, and a cutting-edge transformer-based approach (BERTopic) Devlin et al. (2019); Grootendorst (2020), which requires a novel simulation-based pipeline for robust effect estimation. To achieve this, we treat topic prevalence as a continuous outcome and apply robust regression-based effect estimation techniques. For the STM model, we use the established `estimateEffect` function of the `stm` package in R Roberts, Stewart, and Tingley (2019), which is specifically designed to incorporate

topic-model uncertainty through variational posterior sampling. Recognizing that the BERTopic approach, by its nature, is not a generative probabilistic model and thus lacks native functions for direct effect estimation with integrated uncertainty, we introduce the Covariate Effect Estimation for the BERTopic model (COFFEE) implemented in `Python`. This innovative approach emulates the STM framework by generating multiple bootstrapped topic distributions using BERTopic's `approximate_distribution()` function and subsequently applying Ordinary Least Squares (OLS) regression to estimate the effects of demographic and geographic covariates robustly. While this strategy does not replicate the full probabilistic machinery of STM, it critically enables meaningful methodological comparisons by capturing uncertainty through nonparametric resampling Mooney (1996); Tibshirani and Efron (1993) and aligning the structure of the statistical analysis across distinct modeling paradigms Egger and Yu (2022). This multimodel comparison is crucial for assessing how different underlying modeling assumptions affect both the topic coherence and the interpretability and robustness of subsequent downstream analyses.

By rigorously mapping the connections between research content, researcher demographics, and geography within the Canadian scientific landscape, this study seeks to identify emerging patterns, highlight potential disparities, and provide a solid foundation for developing more equitable, effective, and impactful funding strategies. This work is especially timely and relevant, directly aligning with national initiatives such as the Tri-Agency Equity, Diversity, and Inclusion (EDI) Action Plan *Tri-Agency Equity, Diversity, and Inclusion Action Plan* (2021), which underscores a commitment to fostering a more inclusive and representative research ecosystem across Canada.

The primary objective of this thesis is to explore the evolving landscape of Canadian science, technology, engineering, and mathematics (STEM) funded research through the lens of topic modeling and covariate effect estimation, focusing on a comprehensive analysis of 18 years (2005-2022) of publicly available research proposals funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). Specifically, the objectives are to:

(1) Identify and characterize latent research themes and their evolution within the NSERC funding landscape from 2005 to 2022, and how have these themes evolved over time?

(2) Comparatively evaluate the performance of Latent Dirichlet Allocation (LDA), Structural

Topic Modeling (STM), and BERTopic in uncovering these themes, assessing their coherence, uniqueness, and diversity.

(3) Quantify the influence of researcher gender and geographical location on the prevalence of identified research topics, using both `estimateEffect` for STM and a novel simulation-based approach (COFFEE) for BERTopic .

Methodologically, this work serves as a testament to how modern topic modeling tools, which encompass both established probabilistic models and cutting-edge embedding-based approaches, can be innovatively applied to large-scale textual data to inform science policy and significantly enhance our understanding of the complex research ecosystem. In doing so, we contribute both substantively and methodologically to ongoing efforts to foster a more inclusive, innovative, and forward-looking scientific community in Canada.

The remainder of this thesis is organized as follows: Chapter 2 reviews the relevant research work; Chapter 3 discusses the data used in this study; Chapter 4 and 5 discusses methodologies and main findings of the first and second research objectives, respectively; Chapter 6 concludes and discusses the thesis; and Chapter 7 presents the limitations of this study and draws some directions for future research.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the scholarly literature relevant to understanding research trends and funding dynamics in science and technology. We first examine the evolving landscape of science policy and the role of large-scale textual analysis in this domain. Subsequently, we delve into the theoretical foundations and advancements in topic modeling, followed by a detailed discussion of established research on gender and geographical disparities in scientific funding. Finally, we explore methodological considerations for quantitative inference from textual data.

Investing in research and development is a critical driver of innovation and economic sustainability Ebadi et al. (2020); Stephan (2015). Understanding the downstream impact of grant funding on scientific productivity is crucial for optimizing resource allocation and maximizing societal benefits Ebadi, Zahedi, Jowkar, and Zare (2016); Jacob and Lefgren (2011). This necessitates a comprehensive understanding of evolving research trends for effective and equitable allocation of resources. Recent analyses of global R&D expenditure highlight the increasing prioritization of science and technology as tools for addressing societal challenges, though disparities in funding distribution persist across regions and demographics Ebadi et al. (2020). In Canada, initiatives like the Tri-Agency Equity, Diversity, and Inclusion (EDI) Action Plan *Tri-Agency Equity, Diversity, and Inclusion Action Plan* (2021) reflect a growing recognition of systemic inequities in research ecosystems, underscoring the need for data-driven assessments of funding patterns.

2.1 The Evolving Landscape of Science Policy and Funding Mechanisms

The governance of science and technology, traditionally framed around the linear model of innovation, has evolved to encompass complex interactions between public funding bodies, private industry, and academia Macnaghten (2022). Contemporary science policy aims not only at generating new knowledge but also at fostering societal impact, economic growth, and addressing grand challenges such as climate change and public health crises Burgelman, Chloupková, and Wobbe (2014); Omenn (2006). This necessitates a nuanced understanding of how research ecosystems operate, how funding decisions are made, and their downstream effects on scientific output and societal benefit Smith, Schäfer, and Bernstein (2024). Different national contexts adopt varied approaches, from highly centralized funding agencies to distributed systems, each with unique implications for research priorities and researcher behavior Zhou, Cai, and Lyu (2020). The Canadian context, exemplified by NSERC, represents a model that balances competitive funding with national strategic priorities, including a strong emphasis on equity, diversity, and inclusion Natural Sciences and Engineering Research Council of Canada (2022). Therefore, empirical studies dissecting the dynamics of research funding are critical for evidence-based policy making, allowing agencies to optimize resource allocation and ensure accountability for public investment.

2.2 The Role of Large-Scale Textual Analysis in Science Studies

The proliferation of digital textual data from scientific endeavors, including grant proposals, publications, and patents, has opened new avenues for science policy research. Traditional bibliometric approaches, while valuable for analyzing citation networks and publication patterns, often fall short in capturing the thematic content and conceptual evolution of research fields at scale Chen et al. (2024). This gap has been increasingly filled by computational text analysis methods, particularly topic modeling, which enable researchers to move beyond simple keyword counts to identify latent intellectual structures and emerging research fronts Hankar, Kasri, and Beni-Hssane (2025).

The application of these methods provides a macro-level perspective on scientific activity, allowing for the mapping of interdisciplinary connections, tracking research trajectories, and identifying under-explored or over-saturated areas within a scientific domain. Moreover, when combined with metadata, these techniques offer powerful tools to investigate the socio-technical dimensions of science, such as the influence of demographic characteristics or institutional affiliations on research themes and outcomes Schulze, Wiegrebe, Thurner, Heumann, and Aßenmacher (2024).

2.3 Evolution of Topic Modeling Paradigms

Topic modeling has emerged as a cornerstone of computational text analysis, providing a statistical framework for discovering abstract 'topics' that occur in a collection of documents. Early approaches, such as Latent Semantic Analysis (LSA) Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), used singular value decomposition to identify underlying semantic relationships. However, the introduction of Latent Dirichlet Allocation (LDA) marked a significant paradigm shift. LDA, a generative probabilistic model, posits that each document is a mixture of various topics, and each topic is characterized by a distribution over words Blei et al. (2003). Its Bayesian nature provides a robust framework for inference, making it highly influential in diverse fields. Despite its widespread adoption, LDA operates under the assumption that documents and topics are exchangeable, meaning it does not intrinsically account for metadata or temporal dynamics, which can limit its utility in complex socio-scientific analyses Vayansky and Kumar (2020).

To address these limitations, extensions to LDA have been developed. Hierarchical LDA (hLDA) Griffiths, Jordan, Tenenbaum, and Blei (2003) introduced a hierarchical structure to topics, allowing for the discovery of topics at different levels of granularity. Dynamic Topic Models (DTM) Blei and Lafferty (2006) enabled the study of topic evolution over time, crucial for understanding shifts in scientific research. Structural Topic Models (STM) Roberts et al. (2013) further advanced the field by explicitly incorporating document-level covariates into the generative process, allowing researchers to estimate how metadata—such as author demographics or publication year—influences both topic prevalence and word usage within topics. This integration makes STM particularly powerful for social science applications, providing a more principled approach to exploring the relationships

between textual content and contextual variables.

The most recent wave of innovation in topic modeling has been driven by advancements in deep learning, specifically transformer architectures. Traditional topic models like LDA and STM rely on bag-of-words representations, which discard word order and contextual meaning. Transformer-based models, exemplified by BERT Devlin et al. (2019), generate contextualized embeddings, meaning that the vector representation of a word changes based on its surrounding words. This capability allows for a far richer semantic understanding of text. BERTopic Grootendorst (2020) leverages these contextual embeddings by combining them with dimensionality reduction techniques like UMAP Healy and McInnes (2024) and clustering algorithms such as HDBSCAN McInnes, Healy, and Astels (2017) to identify coherent topics. This approach offers enhanced interpretability and the ability to capture more nuanced thematic patterns compared to traditional methods, especially in highly specialized domains. However, its non-probabilistic nature presents unique challenges for statistical inference, requiring alternative approaches for uncertainty quantification and covariate effect estimation, challenges that contemporary research, including this thesis, aims to address through novel methodological pipelines.

2.3.1 Metrics for Assessing Topic Quality and Interpretability

The evaluation of topic models extends beyond merely identifying clusters of words; it critically involves assessing the quality and interpretability of the discovered topics. A topic model is most useful when its output is not only statistically sound but also semantically meaningful to human experts Mimno, Wallach, Talley, Leenders, and McCallum (2011). This emphasis on interpretability has led to the development of various quantitative metrics that aim to approximate human judgment of topic quality.

Topic Coherence, as measured by C_v, is one of the most widely adopted metrics. It quantifies the semantic relatedness of words within a topic by evaluating the co-occurrence patterns of its top words within the original corpus or a relevant external corpus Röder, Both, and Hinneburg (2015). A high coherence score indicates that the words frequently appear together, suggesting a semantically consistent and thus more interpretable topic. Growing evidence indicates that contextualized embedding-based models tend to produce more human-understandable topics due to their ability to

group semantically similar terms more effectively.

Beyond coherence, metrics like Topic Uniqueness and Topic Diversity address other critical aspects of model performance. Topic uniqueness, which quantifies the distinctiveness of top words across different topics within a model, ensures that topics are not merely repeating the same set of common words. High uniqueness suggests that each topic captures a distinct conceptual space. Similarly, Topic Diversity, measured by the proportion of unique words among all top words across topics, reflects the breadth of vocabulary used to define the thematic landscape. A high diversity score implies that the model is generating a rich and varied set of themes, rather than collapsing disparate concepts. These metrics collectively provide a quantitative lens through which to compare how different models resolve the thematic structure of a corpus, offering valuable insights into their comparative strengths and weaknesses in capturing semantic content.

While topic modeling is increasingly used to analyze research trends, the role of demographic factors, such as gender, in shaping research priorities remains relatively less explored. Studies have consistently demonstrated gender disparities within scientific research Abramo, D'Angelo, and Murgia (2013); van den Besselaar and Mom (2022). For instance, Larivière et al. Larivière, Ni, Gingras, Cronin, and Sugimoto (2013) highlighted global gender disparities, indicating that women are often underrepresented in scientific authorship. Similarly, van Arensbergen et al. Van Arensbergen, Van der Weijden, and Van den Besselaar (2012) found that gender differences in scientific productivity and funding success are a persistent phenomenon across various disciplines, emphasizing the continued need for more inclusive approaches to funding allocation. In Canada, persistent gender gaps in Canadian STEM funding, particularly in engineering and physical sciences, have been documented, showing, for instance, higher rejection rates for early-career women scientists and slightly less funding for successful female applicants Ogden (2019). Hajibabaei et al. Hajibabaei, Schiffauerova, and Ebadi (2022, 2023) found that in AI research networks, while scientific performance is key for both genders to achieve central roles, women face subtle disadvantages in becoming "local influencers", which further highlights ongoing gender disparities. These disparities are potentially exacerbated by biases in evaluation processes, as evidenced by studies suggesting that grant applications led by women may be evaluated less favorably than those led by men, even when scientific quality is comparable Bornmann, Mutz, and Daniel (2007b); Wennerås and Wold

(1997); Witteman et al. (2019).

2.3.2 Gender Disparities in Research and Funding: Disciplinary Segregation and Systemic Biases

The literature consistently highlights that gender disparities in science are not uniformly distributed across disciplines but often manifest as a form of 'horizontal segregation,' where women tend to be overrepresented in specific fields (e.g., life sciences, social sciences, humanities) and underrepresented in others (e.g., engineering, computer science, physics) Ecklund, Lincoln, and Tansey (2012); Jaramillo, Macedo, Oliveira, Karimi, and Menezes (2025).

The persistence of gender disparities in research funding is a complex issue, often attributed to a confluence of factors beyond simple meritocratic evaluation. Studies have highlighted the role of implicit bias in peer review processes, where evaluators, irrespective of their own gender, may unconsciously favor proposals from male applicants or those aligned with traditionally male-dominated research areas Bornmann et al. (2007a); Schmaling and Gallo (2023). This bias can manifest in subtle ways, such as differences in language used in recommendation letters or perceptions of competence and leadership potential Schmader, Whitehead, and Wysocki (2007). Beyond individual biases, structural factors contribute significantly. These include historical under-representation of women in senior academic positions, which affects mentorship opportunities and network building crucial for grant success Deanna et al. (2022). Furthermore, career interruptions, often linked to caregiving responsibilities, disproportionately impact women's publication records and grant application frequency, leading to cumulative disadvantage over time, a phenomenon sometimes referred to as the 'leaky pipeline' Sato, Gygax, Randall, and Schmid Mast (2021).

Policy interventions, such as mandated gender quotas on review panels, unconscious bias training, and specific funding initiatives targeting underrepresented groups, have been introduced globally to mitigate these biases and foster a more equitable research landscape Torres, Collins, Hertz, and Liukkonen (2024). Furthermore, the 'Matthew Effect' Merton (1968) can amplify these initial disadvantages, as researchers who receive less funding early in their careers may find it harder to build the track record necessary for future success. While these studies establish the existence and contributing factors of gender disparities, a detailed, large-scale, and comparative analysis of their

influence on specific research themes within grant proposals over time, particularly in the Canadian context, remains an important area for further empirical investigation.

Geographical disparities in research funding allocation persist as a significant challenge, reflecting demographic inequities. Breschi and Lissoni Breschi et al. (2010) illustrate how regional proximity and inventor mobility drive knowledge spillovers, emphasizing the role of geography in shaping innovation. The concentration of funding in metropolitan hubs can marginalize researchers in peripheral regions, limiting access to resources and collaboration opportunities Asheim et al. (2016); Rodríguez-Pose (2018). This spatial inequity aligns with the "Matthew Effect" in science, where established centers attract disproportionate resources, exacerbating regional imbalances. Broader studies on regional innovation systems Asheim et al. (2016) and the socio-economic impacts of regional inequality Rodríguez-Pose (2018) highlight the importance of considering geographical factors in science policy.

2.3.3 Geographical Disparities and Regional Innovation Systems in Research Funding

Geographical disparities in research funding and output are deeply intertwined with theories of regional innovation systems and the spatial dynamics of knowledge creation. The Matthew Effect, where initial advantages lead to cumulative benefits, is particularly salient in this context, contributing to the concentration of research excellence and funding in a few prominent 'science hubs' or 'knowledge cities' Florida (2002). This phenomenon is often explained by the concept of knowledge spillovers, where the geographical proximity of researchers, institutions, and firms facilitates informal interactions, shared tacit knowledge, and collaborative opportunities, thereby reinforcing existing centers of excellence Boschma (2005). These agglomeration economies can lead to a self-reinforcing cycle, attracting top talent and further investment, while regions outside these hubs struggle to develop competitive research capacities Porter (2008).

Furthermore, the structure of funding mechanisms can inadvertently exacerbate these disparities. Competitive funding, while promoting excellence, can disproportionately benefit well-established institutions with strong grant-writing infrastructure and extensive networks Petersen (2021). Understanding these spatial inequalities is crucial for designing policies that promote more balanced

regional development, foster innovation across a broader geographical spectrum, and ensure that scientific benefits are distributed more equitably across a nation McCann (2001).

2.3.4 Quantitative Inference from Text: Methodological Advances and Challenges

The burgeoning field of 'text as data' has transformed the social sciences, enabling researchers to systematically extract and analyze information from vast unstructured textual corpora to answer substantive questions Grimmer and Stewart (2013). Central to this endeavor is the ability to move beyond descriptive summaries of textual content to formally test hypotheses about how textual features relate to external variables or outcomes. This often involves integrating topic model outputs into traditional statistical frameworks, such as regression analysis, to quantify the influence of various covariates on thematic prevalence or content Glenny, Tuke, Bean, and Mitchell (2019).

While probabilistic topic models like STM Roberts et al. (2013) offer integrated solutions for such inference, by design allowing for uncertainty quantification through their generative process, the rise of powerful, non-generative deep learning models like BERTopic Grootendorst (2020) introduces new methodological considerations. These models, while excelling in semantic representation and topic quality, do not inherently provide a probabilistic framework for statistical inference. Consequently, researchers must employ alternative, often non-parametric, methods to estimate effects and quantify uncertainty robustly. Bootstrapping Tibshirani and Efron (1993) stands out as a widely accepted and versatile resampling technique for this purpose. By repeatedly sampling with replacement from the observed data and re-estimating parameters on each resample, bootstrapping allows for the empirical approximation of sampling distributions, providing robust standard errors and confidence intervals even in the absence of traditional parametric assumptions. This approach is particularly valuable for complex models where analytical derivation of sampling distributions is intractable. Such novel simulation-based pipelines exemplify how modern computational methods can bridge methodological gaps, enabling rigorous comparative statistical inference across diverse topic modeling paradigms and advancing the quantitative study of textual data in science policy research Efron and Narasimhan (2020).

The preceding sections have established the critical importance of understanding science policy and funding mechanisms, underscored the transformative potential of large-scale textual analysis

and advanced topic modeling techniques, and highlighted persistent gender and geographical disparities in research. The identified gaps concerning the thematic impact of these demographic factors on grant funding and the need for robust quantitative inference from non-generative topic models collectively underscore the imperative for the empirical investigation presented in this thesis. By addressing these areas, this research aims to contribute to a more nuanced understanding of funding dynamics and inform evidence-based policies for a more equitable research ecosystem.

Chapter 3

Data

In this study, we utilized a comprehensive dataset of scientific projects and research proposals funded by NSERC. The publicly available dataset covers the period from 2005 to 2022. This publicly available dataset includes key features such as: Application ID, Application Title, First name and last name of the Researcher, Organization, Province, Country, Competition Type, Award Amount, and Proposal Summary. Initially, we had $316,521$ data points. After the preprocessing steps detailed in the methodology chapter, the data was reduced to a final count of $78,863$ proposals. The total award amounts and number of applications per year are illustrated in Fig. 3.1

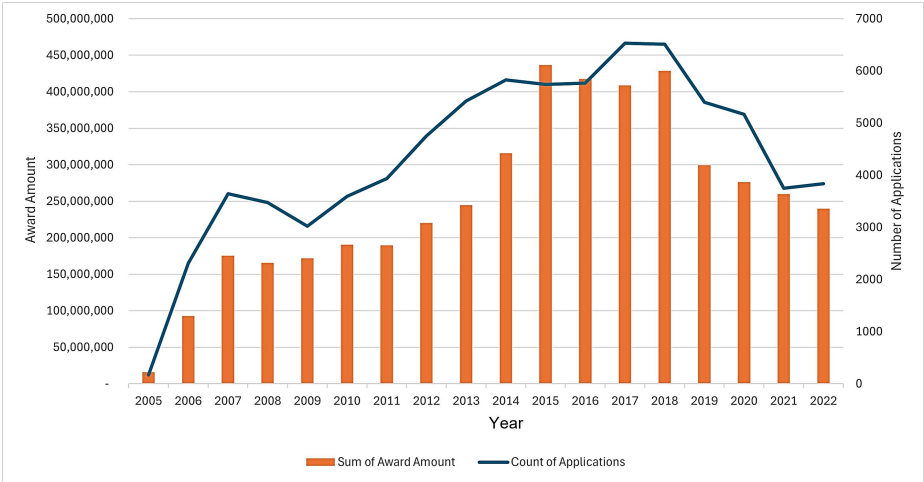

Figure 3.1: Award amounts and number of applications per year

The gender of researchers was determined through a multifaceted approach using the GPT-4

model Achiam et al. (2024). For each researcher, GPT-4 was prompted with their first name, last name, and country of origin, to classify it as “male”, “female”, or “ambiguous”. The prompt was carefully engineered to encourage GPT-4 to utilize its vast training data, including implicit knowledge of linguistic patterns (e.g., gendered name suffixes in various languages) and regional name prevalence. In cases where GPT-4 classified a name as “ambiguous”, we subsequently adopted the most frequently used gender associated with that name historically (e.g., if ”Taylor” was historically predominantly male, it was classified as male). To ensure the reliability of our gender assignments, we conducted a rigorous validation process. A stratified random sample of 500 researchers was manually verified against existing records and common knowledge for accuracy. This verification yielded an overall accuracy rate of 93 percent. Further analysis showed an accuracy of 91 percent for female and 95 percent for male classifications. For names initially classified as “ambiguous” by GPT-4 and subsequently resolved using historical frequency, the accuracy of our classification was verified at 88 percent, which provided confidence in our gender assignments. The resulting gender distribution is shown in Fig 3.2 and 3.3. The chart illustrates the proportions of male and female researchers in the dataset based on the GPT-4 classification and validation process.

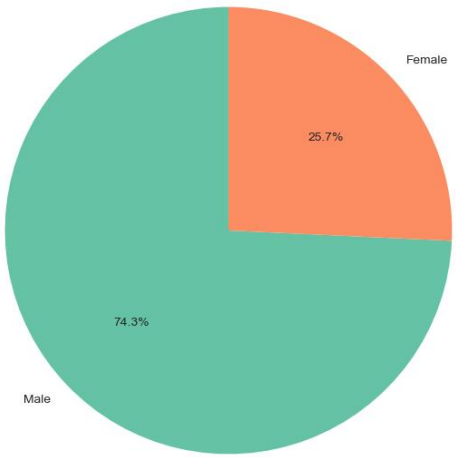

Figure 3.2: Gender Distribution of Researchers

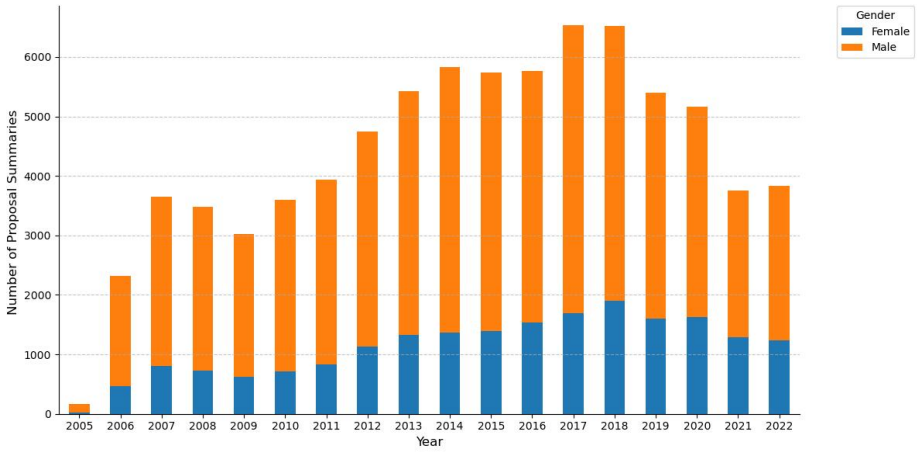

Figure 3.3: Distribution of genders over time by male and female.

Chapter 4

Methodology

Our methodology consists of four key components, as shown in Fig. 4.1: (1) Data collection and pre-processing, (2) Topic modelling, (3) Comparative analysis between models and evaluation, and (4) Effect estimation of geographical locations and gender. The following subsections describe each of these key components in more detail.

4.1 Data preprocessing

The collected data underwent a thorough cleaning and preprocessing process to ensure its compatibility with our modelling techniques. The following steps were implemented:

(1) **Removal of incomplete entries:** All entries lacking application summaries were removed to ensure that each record contained sufficient information for analysis.

(2) **Elimination of duplicates:** Duplicate records were identified and eliminated to maintain data integrity and prevent redundancy.

(3) **Translation of non-English content:** Non-English application summaries (e.g., French, Italian) were translated into English using the `deep_translator` package in Python. To ensure the quality of the translations, a random sample of 5 percent of the translated content was manually reviewed. This verification step confirmed the accuracy of the translations for

subsequent analysis. The original language distribution of the proposals is shown in Fig. 4.2 and 4.3.

(4) **Cleaning of non-textual elements:** Non-textual elements, including punctuation, numbers, and special characters, were removed to streamline the text for processing.

(5) **Text normalization:** The text was converted to lowercase and tokenized, facilitating consistent analysis across all entries.

(6) **Stop-word removal:** Common English stop-words were removed to reduce noise and focus on meaningful content.

(7) **Exclusion of domain-specific terms:** In addition to general stop-words, domain-specific terms such as “Canada”, “NSERC”, and “Research” were excluded to prevent skewing of thematic analysis.

(8) **Lemmatization:** Words were reduced to their root forms by lemmatization, aiding in the consolidation of similar terms.

(9) **Consideration of n-grams:** Bigrams and trigrams were also considered to capture contextually relevant phrases and enhance thematic richness.

This comprehensive preprocessing approach ensured that the dataset was optimized for subsequent modelling and analysis, allowing for accurate exploration of research themes.

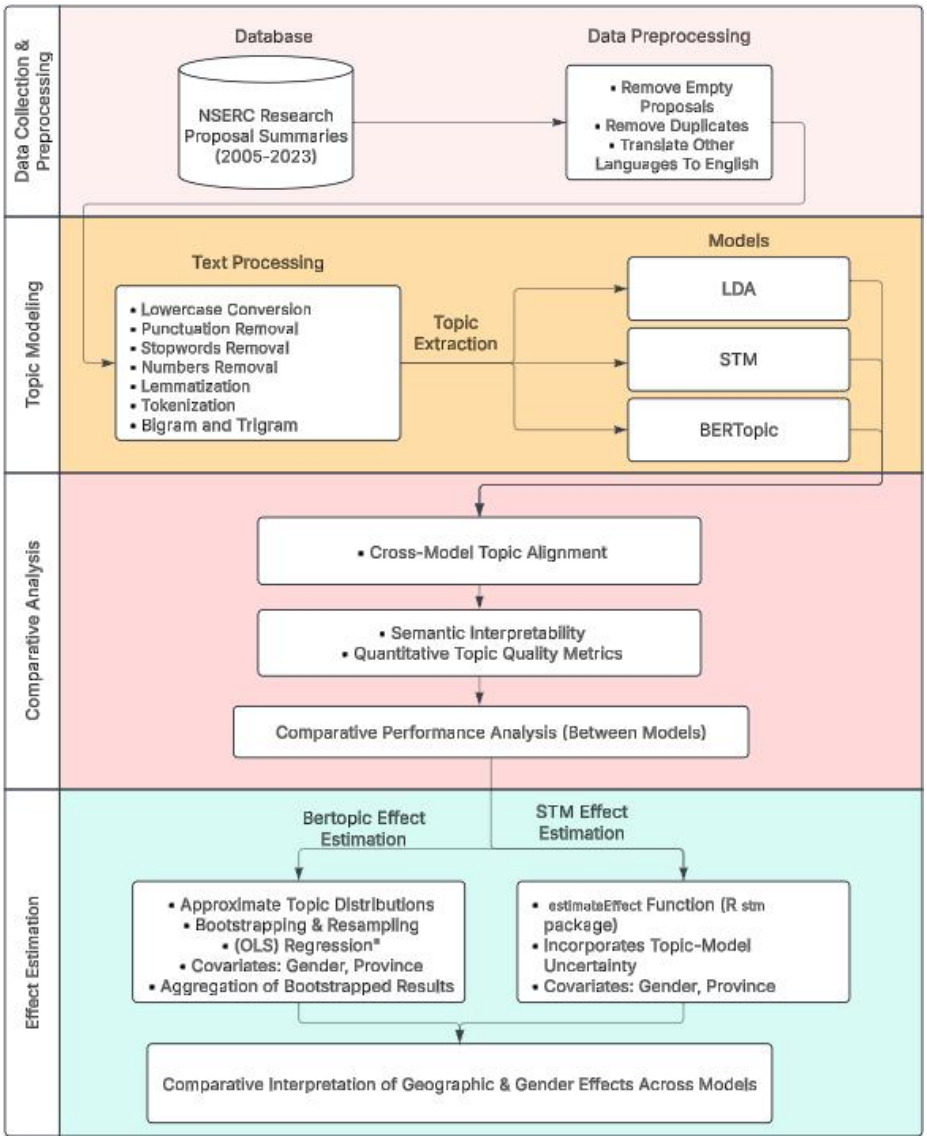

Figure 4.1: High-level flow of the analysis

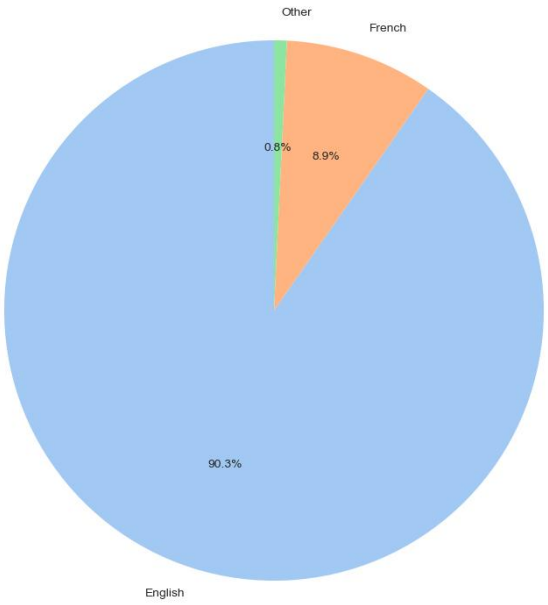

Figure 4.2: Distribution of Proposal Languages. All languages other than English and French are grouped into the 'Other' category.

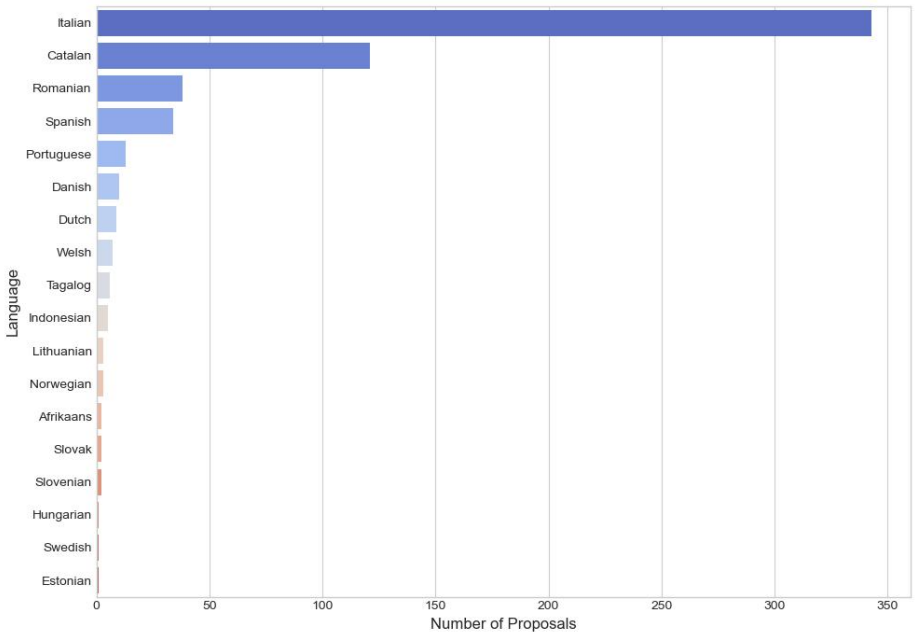

Figure 4.3: Distribution of Proposal Languages other than English and French.

4.2 Topic Modeling

In this thesis, we employed and compared three topic modelling techniques to extract themes from our corpus. Each model is described below.

4.2.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003), is a generative probabilistic model designed to uncover latent thematic structures within a corpus of documents. The fundamental assumption of LDA is that each document can be represented as a mixture of topics, and each topic is characterized by a probability distribution over words. In this framework, documents are generated through a hierarchical probabilistic process: for each document, a distribution over topics is first sampled, and subsequently, for each word position, a topic is drawn from this distribution, followed by a word drawn from the corresponding topic–word distribution.

This process is governed by two Dirichlet priors. The first parameter, parameterized by α, controls the *document–topic distributions* (θ), thereby influencing how many topics are likely to appear within a single document. A smaller value of α encourages documents to be dominated by a small number of topics, while a larger value results in documents exhibiting a broader mixture of topics. The second prior, parameterized by β, controls the *topic–word distributions* (ϕ), which determine the lexical composition of each topic. A smaller β leads to topics that are more focused and defined by a few characteristic words, whereas a larger β produces more diffuse topics covering a wider vocabulary.

The task of inference in LDA is to reverse this generative process and estimate the hidden variables θ and ϕ from the observed words in the corpus. To determine the optimal number of topics (K), coherence scores (C_v) were computed across a candidate range ($2 \leq K \leq 20$) in combination with a qualitative examination of the top-ranked keywords associated with each topic. Based on this procedure, the optimal number of topics was selected as $K = 11$.

4.2.2 Structural Topic Modeling (STM)

The Structural Topic Model (STM), proposed by Roberts et al. (2013), extends traditional topic modeling by incorporating document-level metadata into the estimation of topics. Unlike Latent Dirichlet Allocation (LDA), which treats documents independently of external information, STM models both *topic prevalence* (the proportion of topics within a document, θ_d) and *topic content* (the distribution of words within a topic, ϕ_k) as functions of observed covariates. This formulation enables researchers to test hypotheses about how document characteristics such as publication date, author identity, or geographic region affect the frequency and linguistic structure of topics. For instance, STM can capture whether a topic becomes more prominent during a particular time period or whether its vocabulary shifts across different author groups.

To determine the optimal number of topics (K), we employed the `searchK()` function from the `stm` package, conducting a grid search over a candidate range of $2 \leq K \leq 20$. Based on this evaluation, which considered both quantitative metrics and topic interpretability, the optimal number of topics for the STM was determined to be $K = 11$.

4.2.3 BERTopic

We employed the BERTopic model developed by Grootendorst (2020), which integrates transformer-based embeddings with clustering and class-based term weighting to generate interpretable topics, unlike probabilistic models such as LDA or STM, which rely on word co-occurrence patterns, BERTopic leverages contextual embeddings from Bidirectional Encoder Representations from Transformers (BERT) to capture semantic relationships between words and documents. This enables more nuanced representations of topics, particularly in domains with rich linguistic variability.

The BERTopic pipeline consists of four main stages. First, document embeddings were generated using a pre-trained BERT model, producing high-dimensional vector representations that encode semantic and syntactic information. Due to the high dimensionality of these embeddings, we applied Uniform Manifold Approximation and Projection (UMAP) (Healy & McInnes, 2024), a non-linear dimensionality reduction technique that preserves both local and global structure of

the embedding space. Second, the reduced embeddings were clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes et al., 2017). This density-based approach is well-suited for textual data as it identifies clusters of varying shapes and sizes while automatically excluding noise points, thus eliminating the need to predefine the number of clusters K.

Third, to derive interpretable topic representations, we applied class-based Term Frequency–Inverse Document Frequency (c-TF-IDF). For each cluster c, the c-TF-IDF weight of a word w is defined as:

$$\text{c-TF-IDF}(w, c) = \frac{f(w, c)}{\sum_{w' \in V} f(w', c)} \cdot \log \frac{N}{|\{c' \in C : w \in c'\}|},$$

where $f(w, c)$ is the frequency of word w in cluster c, V is the vocabulary, N is the total number of clusters, and C is the set of all clusters. This formulation highlights words that are particularly characteristic of a given cluster relative to the others, thereby enhancing topic distinctiveness.

Finally, the resulting topics were evaluated using both quantitative and qualitative measures. We computed topic coherence to assess the semantic consistency of keywords, and we manually inspected top-ranked terms to ensure interpretability and relevance. Based on this procedure, the optimal number of topics for the BERTopic model was determined to be $K = 13$.

4.3 Comparative Analysis of Topic Models

To conduct a robust comparison of the topics generated by the three models—LDA, STM, and BERTopic—we adopted a systematic approach combining cross-model topic alignment, semantic interpretability assessment, and quantitative evaluation metrics.

4.3.1 Cross-Model Topic Alignment

To represent each topic consistently, we derived vector representations from the top keywords of each topic. specifically, we computed the mean vector of Sentence-BERT embeddings Reimers and Gurevych (2019) generated for the top-30 keywords of each topic using HuggingFace's `all-MiniLM-L6-v2` pre-trained model. To quantify thematic correspondence and overlap among different topic sets, we

measured the semantic similarity between their vector representations, which were generated in the previous step. We calculated the **semantic similarity** between any two topic vectors, $\mathbf{v}_i$ and $\mathbf{v}_j$, using the **cosine similarity metric**. Based on these pairwise cosine similarity scores, topics identified by different models were grouped to pinpoint **thematic correspondences**. We set the grouping threshold at **0.82**, a value determined through an iterative process of qualitative assessment. We experimented with various thresholds (e.g., 0.7, 0.75, 0.8, 0.85, 0.9) and manually inspected samples of grouped topics to verify their coherence. We found that a lower threshold tended to group dissimilar topics, introducing noise, while a higher one missed clear thematic overlap. The 0.82 threshold emerged as the **optimal balance**, maximizing meaningful thematic correspondences, while minimizing the inclusion of unrelated topic. This value consistently yielded topic groupings that were qualitatively judged as highly coherent. Once grouped by this method, these topic groups were categorized as follows:

- **Triplet Matches** ($n = 5$)**:** Groups consisting of one topic from each model (BERTopic, STM, LDA), where the cosine similarity between all three pairs met or exceeded the threshold.
- **Semi-Matches** ($n = 6$)**:** Groups consisting of two topics from different models with a cosine similarity equal to or greater than the threshold, which were not part of a "Triplet Match".
- **Unique Topics** ($n = 8$)**:** Topics from a single model that did not achieve a cosine similarity above the threshold with any topic from another model.

Additionally, a t-SNE (t-Distributed Stochastic Neighbour Embedding)(Fig. 5.1) visualization was generated using the Sentence-BERT topic embeddings Maaten and Hinton (2008). This visualization offers a lower-dimensional spatial representation of the topic space, facilitating a qualitative assessment of the clustering and interrelationships between topics across different model. To assign a thematic label to each group, we utilized the GPT-4 model Achiam et al. (2024) to synthesize the core concept from the constituent keywords, and each generated label was then subjected to a final qualitative validation to ensure its fidelity to the underlying concepts of the keyword cluster.

4.3.2 Topic Quality Evaluation

For a quantitative comparison of topic quality, three metrics were computed for topics that formed triplet matches. These metrics were calculated using the original lemmatized corpus and the derived dictionary:

- **Topic Coherence (C_v):** This metric measures the semantic consistency within a topic based on word co-occurrence patterns in the corpus Röder et al. (2015). A higher C_v score indicates greater interpretability and reliability. Formally, for a topic T with top words W_T, coherence is computed as the average pairwise Normalized Pointwise Mutual Information (NPMI) score:

$$C_v(T) = \frac{1}{\binom{|W_T|}{2}} \sum_{\substack{w_i, w_j \in W_T \\ i<j}} \text{NPMI}(w_i, w_j),$$

  where:

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log(P(w_i, w_j)+\epsilon)},$$

  with $P(w_i, w_j)$ and $P(w_i)$ estimated from the corpus using co-occurrence counts, and ϵ a small constant to ensure numerical stability.

- **Topic Uniqueness:** This metric quantifies the distinctiveness of a topic's top words within the set of evaluated triplet topics. It is calculated as the average inverse frequency of each word among all top words across the triplet topics for a given model. A higher score indicates less word overlap and greater distinctiveness. For a topic T with top words W_T:

$$\text{Uniqueness}(T) = \frac{1}{|W_T|} \sum_{w \in W_T} \frac{1}{\text{count}(w, \mathcal{W}_{\text{all}})},$$

  where $\mathcal{W}_{\text{all}}$ is the multiset of all top words from the triplet topics for the model, and $\text{count}(w, \mathcal{W}_{\text{all}})$ denotes the number of occurrences of word w in $\mathcal{W}_{\text{all}}$. The model's uniqueness score is the average of $\text{Uniqueness}(T)$ across all its matched topics.

- **Topic Diversity:** This metric represents the vocabulary range across the evaluated topics for a given model. It is calculated as the proportion of unique words among all top words from

the triplet topics. Formally:

$$\text{Diversity} = \frac{\left|\bigcup_{T \in \mathcal{M}} W_T\right|}{\sum_{T \in \mathcal{M}} |W_T|},$$

where $\mathcal{M}$ is the set of matched triplet topics for the model, and $\sum_{T \in \mathcal{M}} |W_T|$ is the total number of words (with duplicates) across all topics. Higher scores indicate a broader range of distinct terms across topics.

4.4 Covariate Effect Estimation

After extracting topics, we aimed to estimate the effects of various variables on research theme. Specifically, we examined the relationship between the research topics and both **geographic location** and **gender**. Given the uneven distribution of scientific publications across provinces, we preprocessed the geographical location variable. Locations with fewer than 1000 publications were combined into a single "Other" group. This step enhances the statistical stability of the regression estimates for the larger, more frequent provinces while still accounting for the contributions of publications from smaller regions. (Figure: 4.4)

We estimated the effect of province and gender on the prevalence of each topic identified by the STM and BERTopic model.We treated topic prevalence, i.e., the proportion of a document d assigned to a given topic k ($\theta_{d,k}$), as the dependent variable in a regression model, with province and gender as key predictors separately.The resulting categorical variables were treated with sum contrasts in the regression model. Sum contrasts compare the mean of each category to the grand mean of the dependent variable across all observations.

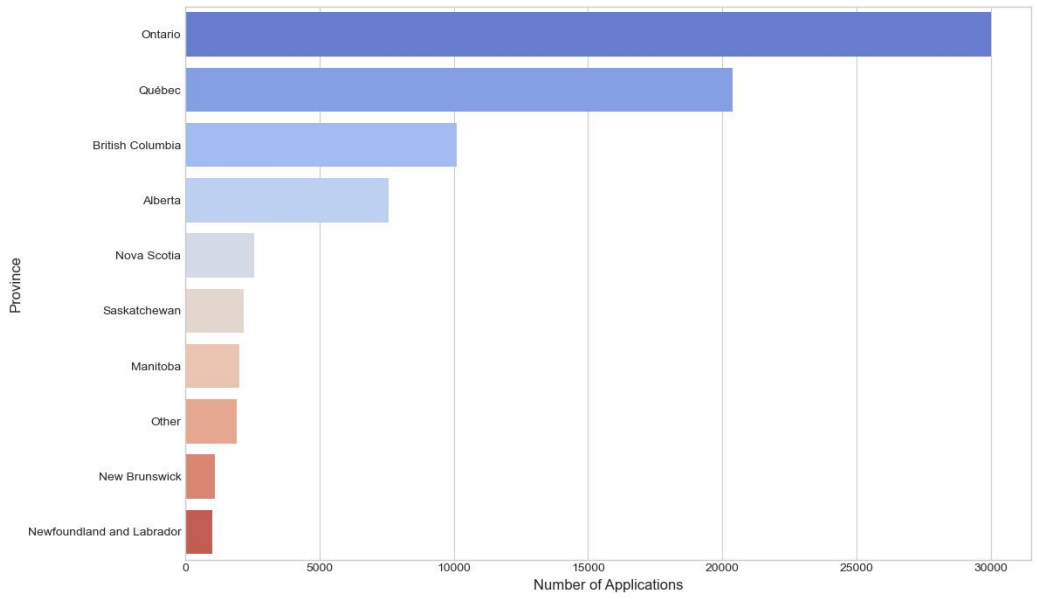

Figure 4.4: Geographical Distribution of Applications by Province. Provinces with fewer than 1000 applications have been grouped into a single 'Other' category.

In this context, the intercept of the regression represents the estimated grand mean topic proportion across all publications, and the coefficient for each province or gender category represents the estimated difference between that category's mean topic proportion and this grand mean. The general form of the regression model for topic k is:

$$\theta_{d,k} = \beta_{0,k} + \sum_{j=1}^{P-1} \beta_{j,k} C_{d,j} + \epsilon_{d,k}$$

where $\theta_{d,k}$ is the proportion of document d in topic k, $\beta_{0,k}$ is the intercept (overall mean), $\beta_{j,k}$ is the coefficient for the j-th category (representing the difference from the overall mean), $C_{d,j}$ are the sum contrast variables for the P categories of the relevant covariate (e.g., gender or geographical location), and $\epsilon_{d,k}$ is the error term.

For the STM model, we utilized the established `estimateEffect` function in the `stm` R package. This function is specifically designed to estimate the effects of document metadata on topic prevalence, incorporating uncertainty by drawing samples from the model's variational posterior distribution. This method provides a principled, model-integrated approach to inference within the

STM framework.

For the BERTopic model, which lacks a native probabilistic structure and therefore does not support the same model-integrated inference as STM, we designed and developed an algorithm, called Covariate Effect Estimation BERTopic (COFFEE). This algorithm conducts a comparable statistical analysis and quantifies uncertainty. Unlike the STM approach, which samples from a model-specific posterior, COFFEE employs bootstrapping, a general resampling technique, to estimate effects. The COFFEE algorithm was developed in Python, providing a robust alternative for effect estimation in the BERTopic framework, as outlined in Algorithm 1.

The COFFEE algorithm enables us to replicate the statistical objective of R's `estimateEffect` function, which regresses topic proportions on metadata using sum contrast. Additionally, COFFEE provides a non-parametric means of approximating the sampling distribution of coefficient estimates through resampling, thereby, quantifying uncertainty in the absence of a model-based posterior. By mirroring the analytical structure of STM's `estimateEffect`, this approach allows for a direct and methodologically consistent comparison of the estimated relationships between the selected covariates (e.g., province or gender) and topic prevalence across both STM and BERTopic outputs.

Algorithm 1 Covariate Effect Estimation for the BERTopic model (COFFEE)

Require: Documents $D = \{d_1, \ldots, d_n\}$, pre-trained BERTopic model $\mathcal{M}$, covariate data C_D (e.g., a DataFrame containing `Gender` or `Geographical Location`), number of bootstrap samples $N = 25$

Ensure: Estimated coefficients $\hat{\beta}_k$, standard errors $SE(\hat{\beta}_k)$, t-values t_k, and p-values p_k for each topic k and each covariate term

Phase 1: Bootstrapped Data Preparation
Initialize `all_theta_samples` and `all_covariate_data_samples` as empty lists
for $s = 1$ to N **do**
    Resample $D^{(s)}$ and $C_D^{(s)}$ with replacement from D and C_D
    $\Theta^{(s)} \leftarrow \mathcal{M}$.`approximate_distribution`$(D^{(s)})$
    Append $\Theta^{(s)}$ to `all_theta_samples`
    Append $C_D^{(s)}$ to `all_covariate_data_samples`
end for
Phase 2: Per-Sample Regression Analysis
for each topic k **do**
    Initialize `coef_samples[`k`]` and `df_resid_samples[`k`]` as empty lists
    for $s = 1$ to N **do**
        $y^{(s)} \leftarrow$ column k of `all_theta_samples[`s`]`
        $X^{(s)} \leftarrow$ `dmatrix`(CovariateFormula, data = `all_covariate_data_samples[`s`]`)
        if OLS fit is feasible **then**
            Fit OLS: $y^{(s)} \sim X^{(s)}$
            Append coefficients to `coef_samples[`k`]`
            Append degrees of freedom to `df_resid_samples[`k`]`
        end if
    end for
end for
Phase 3: Aggregated Inference and Uncertainty Quantification
for each topic k **do**
    if `coef_samples[`k`]` is not empty **then**
        $\hat{\beta}_k \leftarrow$ mean of `coef_samples[`k`]`
        $SE(\hat{\beta}_k) \leftarrow$ standard deviation of `coef_samples[`k`]`
        $df_k \leftarrow$ median of `df_resid_samples[`k`]`
        $t_k \leftarrow \hat{\beta}_k / SE(\hat{\beta}_k)$
        $p_k \leftarrow 2 \cdot$ `t.sf`$(|t_k|, df_k)$
    else
        $\hat{\beta}_k, SE(\hat{\beta}_k), t_k, p_k \leftarrow$ NaN
    end if
end for

Chapter 5

Results

In this section, the results of our comparative topic modeling analysis are presented, which used BERTopic, STM, and LDA to analyze the thematic structure of the corpus. The findings reveal a clear hierarchy of detail and insight, with BERTopic consistently providing a more nuanced and granular view of the data. We begin with a qualitative and quantitative comparison of the topics, highlighting the unique strengths of each model. This is followed by a detailed discussion of the quantitative evaluation metrics, such as topic coherence and diversity, which support our qualitative observations. Finally, we present the results of our Covariate Effect Estimation (COFFEE) algorithm, which leverages the outputs of BERTopic and STM to statistically link our identified topics with demographic and geographical factors, thereby demonstrating the practical, applied value of our chosen methodology.

5.1 Comparative Analysis of Topic Models

The comparative analysis highlights both similarities and differences in the thematic structures identified by BERTopic, STM, and LDA, providing insights into the strengths and weaknesses of each model. Figure 5.1 visualizes the t-SNE plots of the topics identified by these models. Clustered points indicate semantically similar topics regardless of their originating model, and more isolated points signify distinct thematic spaces. This visual evidence not only confirms quantitative similarities and differences but also serves as a foundational reference for understanding the

thematic relationships discussed in detail below. The clear groupings in the plot correspond to the triple alignments, where BERTopic, STM, and LDA often cluster together. For example, B0, S10, and L5 are grouped near the top-left, representing “Environmental Science & Industrial Processes”, while B1, S2, and L10 are clustered in the lower-right for “Computer Science & Artificial Intelligence”. Conversely, the dispersion of unique topics further from these clusters (e.g., B8, B10, B11, B12 at the periphery) visually indicates their distinct semantic spaces and highlights BERTopic’s tendency to identify more niche and specialized themes. In this section, we delve deeper into these topic structures, providing a detailed exploration of the thematic alignments and distinctions that characterize each model’s outputs.

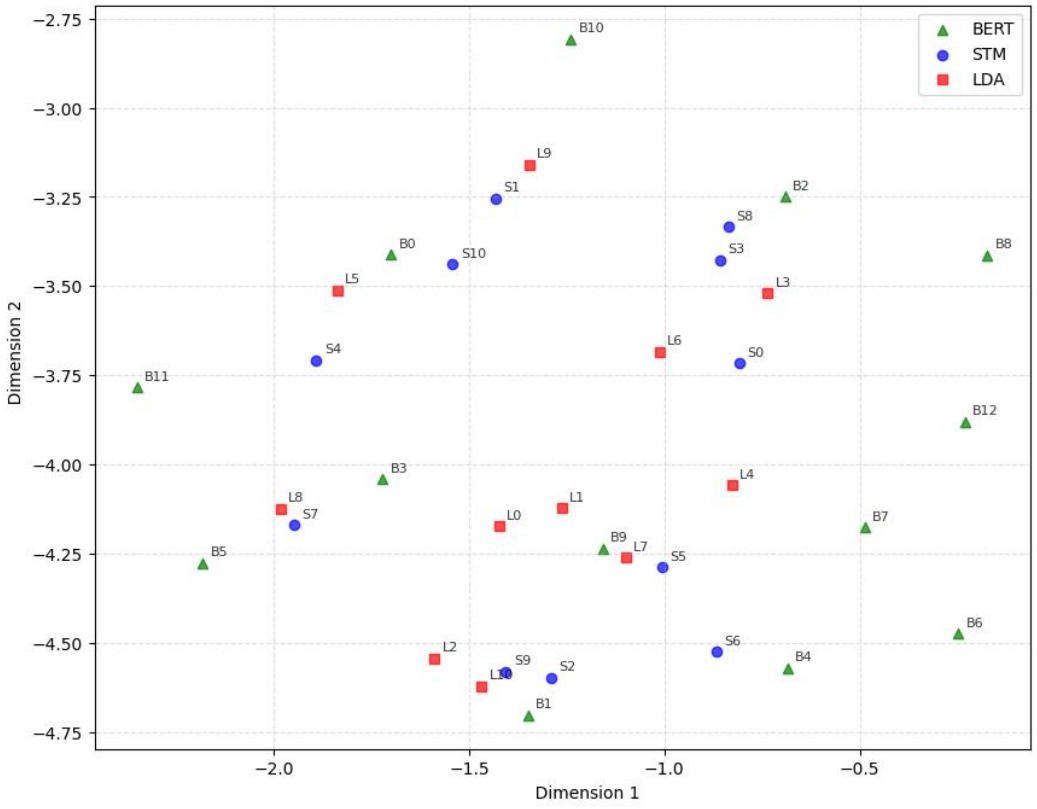

Figure 5.1: t-SNE plot of topics identified by BERTopic, STM, and LDA

5.1.1 Unique Topics

Table 5.1 presents a detailed comparison of unique topics identified by each model, highlighting themes predominantly discerned by a single model. These unique topics are visually corroborated by Figure 5.1, where they often appear as more isolated points, signifying their semantic distinction from the main clusters. Notably, BERTopic stands out with its ability to generate a greater number of specific and unique topics, illustrating its strength in providing a richer semantic understanding of the corpus.

Table 5.1: Comparison of unique topics across models. Numbers in parentheses represent the topic index.

Label	**BERTopic**	**STM**	**LDA**
Musculoskeletal Biomechanics	(6) bone, injury, joint, knee, tissue	-	-
Dental & Oral Biomaterials	(7) dental, implant, tooth, bone, enamel	-	-
Public Health & Vaccine Communication	(8) vaccine, vaccine confidence, education, mental health, kmaw	-	-
Food Science & Agricultural Products	(10) flaxseed, food, fatty acid, mustard, ingredient	-	-
Advanced Energy Storage & Electronics	(11) battery, energy storage, ion battery, solar cell, transistor	-	-
Organ Transplantation & Biofabrication	(12) organ, kidney, transplant, donor, blood	-	-
Plant Pathology & Crop Genetics	-	(3) disease, plant, dna, gene, bacterium	-
Theoretical & Computational Sciences	-	(9) theory, flow, dynamic, simulation, computational	-

BERTopic's distinctive strength is exemplified by its unique topic "Public Health & Vaccine Communication" (B8) topic. Its top ten keywords, `vaccine`, `vaccine confidence`, `education`, `mental health`, `Mi kmaw`, `kmaw`, `covid 19`, `19`, `social media`, `indigenous`, reveal a particular and timely theme, encompassing not just medical terms but also sociopolitical dimensions such as `vaccine confidence` and culturally specific references (`Mi'kmaw`, `indigenous`).

This level of fine-grained resolution, capable of capturing nuanced, multi-word expressions and contextual relevance, was not observed in the topics generated by STM or LDA, which would likely have subsumed these terms into broader health or social topics. This illustrates BERTopic's capability in generating more specific and insightful topics by discerning specialized, emerging, or highly sensitive research areas that other models might overlook due to their reliance on less context-aware word representation. Similarly, BERTopic's unique topic, "Organ Transplantation & Biofabrication" (B12), with keywords such as {`organ, kidney, transplant, transplantation, donor`}, and other keyterms such as `bioink` and `bioprinting`, underscores its ability to isolate specialized subfields within medicine. This capability to delve into niche scientific domains highlights BERTopic's enhanced sensitivity to context and specificity.

In contrast, STM also produced meaningful unique topics, such as "Plant Pathology & Crop Genetics" (S3) with top keywords such as {`disease, plant, crop, dna, gene, bacterium`}. However, these topics often exhibited a broader scope compared to BERTopic's highly specialized themes. This may suggest that while STM is effective at identifying significant research areas, it may not capture the same level of detail in emerging or specialized topics as BERTopic.

In this analysis, LDA did not produce any unique topics. All LDA topics had at least a semi-match, i.e., partial thematic alignment, with those from BERTopic or STM. This pattern suggests that LDA's reliance on word co-occurrence may lead to the identification of more generalized topics, often blending distinct concepts rather than distinguishing them. Consequently, LDA provides a less detailed thematic landscape compared to the more context-aware capabilities of BERTopic.

5.1.2 Triple Alignment

Five topics were consistently identified by all three topic models, indicating a consistent agreement across BERTopic, STM, and LDA on several prominent research themes within the corpus. This degree of alignment, evidenced by an average cosine similarity of 0.935, serves as a validation point for the inherent salience of these topics in the dataset, regardless of the underlying algorithmic approach. Although the models fundamentally agreed on the core subject matter for these themes, a closer examination of their respective top-10 keywords reveals subtle but essential differences in the emphasis or granularity of the topics, highlighting the unique lens each model applies.

Table 5.2: Consensus topics across BERTopic, STM, and LDA. Numbers in parentheses represent the topic index.

Label	BERTopic	STM	LDA	Avg. Sim.
Environmental Science & Industrial Processes	(0) water, energy, climate, gas, soil	(10) water, production, oil, treatment, gas	(5) chemical, reaction, gas, organic, compound	0.935
Computer Science & Artificial Intelligence	(1) network, algorithm, software, communication, wireless	(2) network, software, technology, algorithm, communication	(10) network, computer, software, theory, algorithm	0.925
Neuroscience & Cognitive Science	(4) brain, memory, neuron, neural, mechanism	(6) learn, brain, memory, neural, visual	(4) cell, tissue, brain, signal, memory	0.915
Molecular Biology & Biotechnology	(2) protein, gene, plant, genetic, disease	(0) cell, protein, mechanism, molecular, role	(6) protein, molecular, mechanism, molecule, biological	0.892
Materials Science & Applied Physics (Imaging & Photonics)	(3) polymer, imaging, laser, optical, molecular	(7) quantum, material, light, optical, particle	(2) image, optical, sensor, measurement, light	0.867

For the "Environmental Science & Industrial Processes" domain, all three models delineated a coherent topic centred on natural resources, energy, and industrial operations. BERTopic's B0 topic connected industrial activities with broader environmental concerns, using keywords such as {`water, energy, climate, gas, soil, technology, industry, carbon, oil, production`}. In contrast, LDA's L5 topic focused more on chemical processes, with keywords such as {`chemical, reaction, gas, production, organic, compound, metal, water, fuel, produce`}, offering a detailed view of the industry's material aspects. STM's S10 topic's keywords ({`water, production, oil, treatment, company, gas, quality, chemical, industry, environmental`})provided a balanced perspective, integrating industrial operations and resource management, highlighting `treatment` and `quality`.

Another example is the "Computer Science & Artificial Intelligence". Here, BERTopic's B1 topic, `datum`, `network`, `algorithm`, `software`, `user`, `service`, `communication`, `wireless`,

technology, computer swiftly introduced foundational computing term. Yet, when considering its broader set of associated keywords, terms such as learning and machine appear, indicating the capture of the AI and machine learning paradigm. This suggests that BERTopic's contextual understanding allowed it to signal a more advanced and contemporary focus within computer science. Conversely, LDA's L10 keywords ({network, computer, software, user, theory, algorithm, object, communication, space, computational}) clearly defined the core elements of computer science. Still, they lacked the immediate semantic cues for recent AI-related concepts. On the other hand, STM's S2 topic ({datum, network, software, technology, service, algorithm, user, communication, challenge, industry}) provided a solid, applied computer science perspective.

The consistent identification of "Molecular Biology & Biotechnology" also serves as a strong testament to the models' ability to capture well-established scientific field BERTopic's B2 ({protein, gene, specie, plant, genetic, population, disease, animal, food, fish}), STM's S0 ({cell, protein, mechanism, molecular, role, tissue, signal, cellular, regulate, gene}, and LDA's L7 ({protein, molecular, mechanism, molecule, biological, interaction, disease, role, animal, identify}) all yielded highly coherent and directly relevant terms that define this scientific domain. In this specific instance, the remarkable semantic overlap across all three models indicates a robust and clear thematic signal, suggesting that for such well-defined and widely discussed scientific areas, all three models can effectively extract the core concepts with high fidelity.

The quantitative evaluation of topic quality, as summarized in Table 5.3, offers further insights into the comparative performance of the model BERTopic demonstrated the highest C_V (0.638), outperforming STM (0.604) and LDA (0.569), suggesting that BERTopic's embedding approach more effectively groups semantically similar words, resulting in topics that are more interpretable and cohesive. All models showed high average topic uniqueness, with BERTopic slightly leading (0.963), indicating distinct topic definitions due to minimal repetition of words across topic LDA and STM excelled in topic diversity (0.960 and 0.959, respectively), while BERTopic had slightly lower diversity (0.953), reflecting its focus on generating more unique, specialized topics rather than using a broad vocabulary across common theme.

Table 5.3: Quantitative evaluation of topic quality.

Model	Average Coherence (C_V)	Average Uniqueness	Average Diversity
BERTopic	0.638	0.963	0.953
STM	0.604	0.959	0.959
LDA	0.569	0.960	0.960

5.1.3 Partial Alignment

The six partially aligned topics (see Table 5.4) highlight instances where two models identified a similar theme, while the third either did not resolve a comparable topic or subsumed its component words into a different, broader theme. Interestingly, all observed partial alignments were between STM and LDA, or BERTopic and LDA, with no strong $(BERTopic - STM)$ pairings reaching the defined similarity threshold in this analysis. This pattern may suggest that while BERTopic shares some thematic boundaries with LDA, and STM with LDA, the direct semantic alignment between BERTopic and STM's specific topic groupings may be less pronounced, perhaps due to their distinct algorithmic approaches to word and document representation. For example, the strong alignment between LDA's L9 and STM's S1 topics on "Climate & Aquatic Ecology" demonstrated a shared understanding of environmental and ecosystem assessment, both providing relevant terms like {`water, carbon, forest, climate, ecosystem`}. In another instance, BERTopic's B9 and LDA's L1 topics showed significant agreement on "Mechanical Engineering & Sports Technology", with BERTopic's keywords such as {`design, system, control, structure, vehicle, model, sensor, movement, frame`} closely matching LDA's focus on {`system, control, design, vehicle, structure, model, data, engineering, movement`}.

These observed differences between models can be attributed to their underlying algorithms:

- **BERTopic:** Leverages pre-trained transformer models to generate context-aware word embedding. This allows it to capture nuanced semantic relationships and identify more granular, specific, and often novel topics, as evidenced by its higher number of highly specialized, unique topics. Its high coherence also suggests its ability to form semantically tightly-knit

Table 5.4: Partially aligned topics across BERTopic, STM, and LDA. Numbers in parentheses represent the topic index.

Label	BERTopic	STM	LDA	Avg. Sim.
Climate & Aquatic Ecology	-	(1) climate, water, forest, ecosystem, soil	(9) water, climate, soil, fish, ecosystem	0.941
Imaging & Sensing Applications	-	(5) image, sensor, detection, patient, medical	(7) environment, flow, rate, measurement, region	0.890
Energy & Advanced Materials	-	(4) energy, material, fuel, polymer, heat	(0) material, energy, industry, polymer, operation	0.884
Population & Evolutionary Biology	-	(8) species, population, plant, evolution, fish	(3) gene, population, evolution, specie, gene_expression	0.868
Mechanical Eng. & Sports Technology	(9) bike, composite, sport, suspension, ride	-	(1) mechanical, structural, vibration, weight, force	0.852
Quantum & Nuclear Physics	(5) quantum, star, galaxy, matter, physics	-	(8) nuclear, chip, microfluidic, radiation, spectroscopy	0.823

and interpretable clusters of words. This makes BERTopic particularly insightful for uncovering fine-grained thematic structures and emerging areas within a dataset.

- **STM:** A statistical topic model that can incorporate metadata. Its topics often represent coherent semantic units and demonstrate good overlap with both BERTopic and LDA for a broad theme. STM's unique topics are also interpretable, showing a robust capability to identify distinct themes, though generally at a slightly broader level of granularity than BERTopic's most specialized topics.

- **LDA:** A generative probabilistic model that assumes topics are distributions over words and documents are distributions over topics. LDA excels at identifying common themes by observing word co-occurrence patterns, often producing broader, more generalized topics. In this analysis, all LDA topics had at least partially aligned topics with other models, and it demonstrated substantial topic diversity for these broader themes. However, LDA's main comparative weakness lies in its tendency to produce more generalized topics and its inability to resolve the highly specific or novel topics that BERTopic consistently uncovers. This often results in less granular and potentially less actionable insights for specialized research.

5.2 Covariate Effect Estimation

Our research aims to demonstrate the efficacy of BERTopic in identifying intricate relationships between funded research topics and demographic factors, such as geographical location and gender. A key contribution of this work is the development of the Covariate Effect Estimation BERTopic (COFFEE) algorithm, which allows for robust statistical analysis and effect estimation from BERTopic's non-probabilistic outputs. By systematically comparing the insights yielded by BERTopic (processed via COFFEE)(see Tables A.1 and A.3 in Appendix A) against those obtained from Structural Topic Model (STM)(see Tables A.2 and A.4 in Appendix A, for more details), we highlight BERTopic's ability to not only corroborate established patterns but also to reveal unique, granular insights that traditional bag-of-words models like STM may overlook.

5.2.1 Geographical Relationships

BERTopic's analysis revealed distinct regional research specializations, offering a level of precision and insight into provincial contributions.(Figure 5.2 and Figure 5.3)

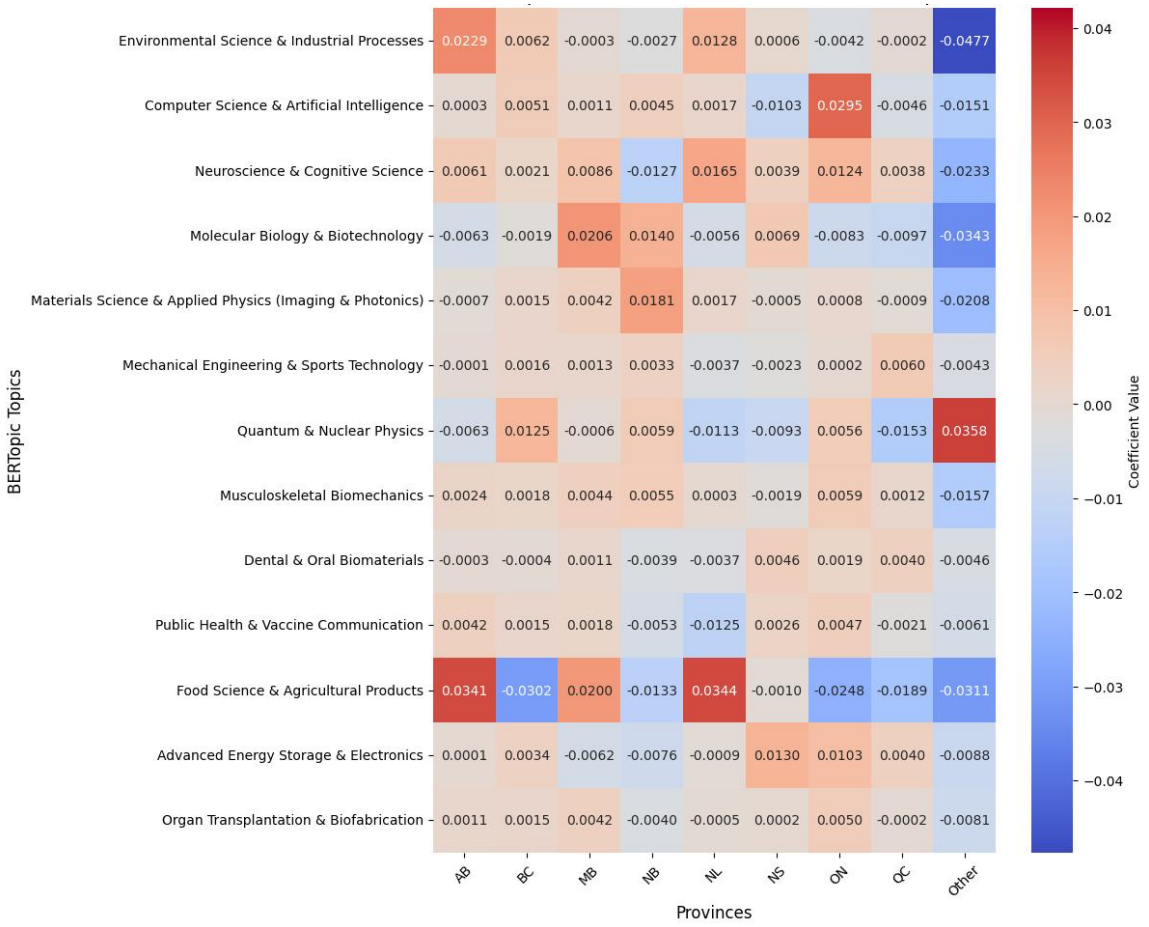

Figure 5.2: Heatmap of BERTopic Topic Provincial Effect Coefficients

For the topic encompassing "Environmental Science & Industrial Processes" (Table 5.5), BERTopic identifies a significant positive effect in Alberta (Estimate: 0.0229, $p < 0.0001$). This finding is strongly corroborated by STM's corresponding "Environmental Science & Industrial Processes" topic (Estimate: 0.0202, $p < 0.0001$), aligning with Alberta's well-established prominence in energy and environmental research Dubé et al. (2021). Moreover, BERTopic uniquely detects a statistically significant positive effect for this topic in Newfoundland and Labrador (Estimate: 0.0128, $p = 0.0224$), a connection linked to its own specific industrial and environmental context Gray

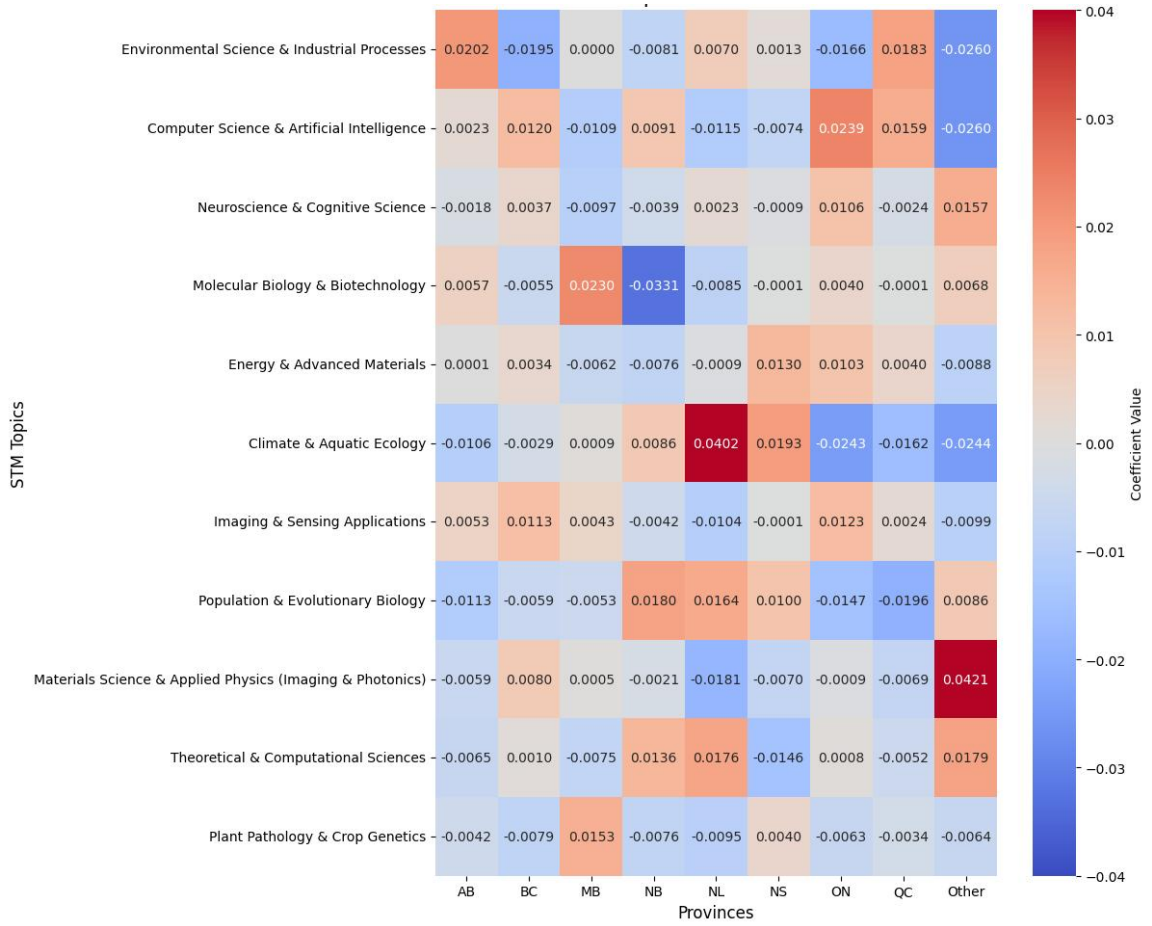

Figure 5.3: Heatmap of STM Topic Provincial Effect Coefficients

(2005). In contrast, STM failed to find a significant effect for this province (Estimate: 0.0070, $p = 0.1952$). This difference highlights the superior granularity of BERTopic. Thanks to its use of contextual embeddings, our COFFEE-enhanced approach can identify cohesive, region-specific research clusters that less semantically rich models, such as STM, may overlook.

Similarly, for the domain of "Computer Science & Artificial Intelligence" (Table 5.6), both BERTopic and STM confirm strong positive estimated effects in Ontario (BERTopic: 0.0295, $p < 0.0001$; STM: 0.0239, $p < 0.0001$) and British Columbia (BERTopic: 0.0051, $p = 0.0256$; STM: 0.0120, $p < 0.0001$). These consistent findings align with the well-recognized status of these provinces as major technology and AI research hubs Arnaout et al. (2024); Wilson (2022), exemplified by institutions like the Vector Institute in Ontario (Vector Institute for Artificial Intelligence,

Table 5.5: BERTopic Provincial Effects: Environmental Science and Industrial Processes

Topic	Province	Estimate	Std. Error	t-value	p-value
Environmental Science & Industrial Processes	Intercept	0.0568	0.0014	39.6711	<0.0001
	Alberta	0.0229	0.0028	8.1083	<0.0001
	British Columbia	0.0062	0.0022	2.8072	0.0050
	Manitoba	−0.0003	0.0038	−0.0818	0.9348
	New Brunswick	−0.0027	0.0057	−0.4749	0.6348
	Newfoundland and Labrador	0.0128	0.0056	2.2831	0.0224
	Nova Scotia	0.0006	0.0039	0.1566	0.8756
	Ontario	−0.0042	0.0021	−2.0046	0.0450
	Other	−0.0477	0.0021	−22.7762	<0.0001
	Québec	−0.0002	0.0021	−0.0934	0.9255

2025). A notable divergence appears in Nova Scotia: BERTopic detects a significant negative estimated effect (Estimate: -0.0103, $p = 0.0024$), whereas STM's effect is non-significant (Estimate: -0.0074, $p = 0.0807$). This difference suggests that BERTopic's heightened sensitivity to regional variations in research focus indicates Nova Scotia's comparatively smaller concentration of high-impact research within this specific AI domain, relative to the tech-centric provinces.

Table 5.6: BERTopic Provincial Effects: Computer Science and Artificial Intelligence

Topic	Province	Estimate	Std. Error	t-value	p-value
Computer Science & Artificial Intelligence	Intercept	0.0449	0.0015	30.8181	<0.0001
	Alberta	0.0003	0.0024	0.1369	0.8911
	British Columbia	0.0051	0.0023	2.2319	0.0256
	Manitoba	0.0011	0.0037	0.3078	0.7583
	New Brunswick	0.0045	0.0066	0.6816	0.4955
	Newfoundland and Labrador	0.0017	0.0051	0.3329	0.7392
	Nova Scotia	−0.0103	0.0034	−3.0367	0.0024
	Ontario	0.0295	0.0018	5.1528	<0.0001
	Other	−0.0151	0.0030	−5.0668	<0.0001
	Québec	−0.0046	0.0015	−3.1233	0.0018

In the field of "Molecular Biology & Biotechnology" (Table 5.7), both models consistently

confirm Manitoba's prominence (BERTopic: 0.0206, $p < 0.0001$; STM: 0.0230, $p < 0.0001$). This alignment validates Manitoba's significant research output in life sciences and biotechnology Canadian Society for Molecular Biosciences (2024).

Table 5.7: BERTopic Provincial Effects: Molecular Biology and Biotechnology

Topic	Province	Estimate	Std. Error	t-value	p-value
Molecular Biology & Biotechnology	Intercept	0.0586	0.0010	56.2270	<0.0001
	Alberta	−0.0063	0.0018	−3.4805	0.0005
	British Columbia	−0.0019	0.0024	−0.7730	0.4395
	Manitoba	0.0206	0.0042	4.8793	<0.0001
	New Brunswick	0.0140	0.0062	2.2621	0.0237
	Newfoundland and Labrador	−0.0056	0.0048	−1.1761	0.2396
	Nova Scotia	0.0069	0.0031	2.2466	0.0247
	Ontario	−0.0083	0.0015	−5.3439	<0.0001
	Other	−0.0343	0.0034	−10.0927	<0.0001
	Québec	−0.0097	0.0015	−6.2402	<0.0001

Moving to "Materials Science & Applied Physics", BERTopic uniquely identifies a significant positive effect in New Brunswick (Estimate: 0.0181, $p < 0.0001$). STM's closest topic, "Quantum Physics", does not capture this, showing a non-significant effect (Estimate: -0.0021, $p = 0.6952$). This unique finding underscores BERTopic's ability to precisely detect niche regional research areas, likely due to its contextual embeddings capturing specialized terms and relationships within materials research and applied physics, such as work at CanmetMATERIALS (Natural Resources Canada, 2024).

in "Public Health & Vaccine Communication" (Table 5.9), BERTopic identifies significant positive effects in Alberta (Estimate: 0.0042, $p = 0.0022$), and Ontario (Estimate: 0.0047, $p < 0.0001$), which are supported by the literature, e.g., Lang, Benham, Atabati, and et al. (2021) and Burney, Donelle, and Kothari (2025), respectively. STM lacks a direct equivalent topic.

Overall, these findings showcase BERTopic's capability, enhanced by the COFFEE algorithm, to capture nuanced regional research patterns with a level of detail and specificity that surpasses traditional methods. This results in a more comprehensive understanding of geographical influences

Table 5.8: BERTopic Provincial Effects: Materials Science and Applied Physics

Topic	Province	Estimate	Std. Error	t-value	p-value
Materials Science & Applied Physics	Intercept	0.0292	0.0007	40.5076	<0.0001
	Alberta	−0.0007	0.0017	−0.3927	0.6945
	British Columbia	0.0015	0.0015	0.9940	0.3202
	Manitoba	0.0042	0.0040	1.0559	0.2910
	New Brunswick	0.0181	0.0035	5.1810	<0.0001
	Newfoundland and Labrador	0.0017	0.0042	0.3983	0.6904
	Nova Scotia	−0.0005	0.0026	−0.1872	0.8515
	Ontario	0.0008	0.0012	0.6583	0.5103
	Other	−0.0208	0.0015	−13.7424	<0.0001
	Québec	−0.0009	0.0013	−0.7276	0.4668

on research themes across Canada.

Table 5.9: BERTopic Provincial Effects: Public Health and Vaccine Communication

Topic	Province	Estimate	Std. Error	t-value	p-value
Public Health & Vaccine Communication	Intercept	0.0154	0.0007	22.1753	<0.0001
	Alberta	0.0042	0.0014	3.0594	0.0022
	British Columbia	0.0015	0.0014	1.0831	0.2788
	Manitoba	0.0018	0.0022	0.7915	0.4286
	New Brunswick	−0.0053	0.0024	−2.1951	0.0282
	Newfoundland and Labrador	−0.0125	0.0013	−9.4260	<0.0001
	Nova Scotia	0.0026	0.0018	1.4691	0.1418
	Ontario	0.0047	0.0010	4.6628	<0.0001
	Other	−0.0061	0.0022	−2.7444	0.0061
	Québec	−0.0021	0.0010	−2.1242	0.0337

5.2.2 Gender Relationship

BERTopic's fine-grained topic resolution, powered by its robust contextual embeddings, uncovers significant gender-based patterns in research topics, often offering deeper insights than STM's broader categorizations.

In "Computer Science & Artificial Intelligence" (Table 5.10), both BERTopic (Estimate: -0.0034, $p < 0.0001$) and STM (Estimate: -0.0165, $p < 0.0001$) consistently indicate a stronger association with male researchers. This robust agreement across models aligns with widely documented gender disparities and the underrepresentation of women in STEM fields, particularly in computing and AI (Hango, 2013). The negative estimates for females highlight a gender gap, calling for targeted initiatives to encourage female participation in these critical areas of technological advancement.

Table 5.10: BERTopic Gender Effects: Computer Science and Artificial Intelligence

Topic	**Gender**	**Estimate**	**Std. Error**	**t-value**	**p-value**
Computer Science & Artificial Intelligence	Intercept	0.0461	0.0008	58.0180	<0.0001
	Female	−0.0034	0.0008	−4.2460	<0.0001

Similarly, for "Public Health & Vaccine Communication" (Table 5.11), BERTopic uniquely identifies a significant positive female effect (Estimate: 0.0029, $p < 0.0001$). STM, as previously noted, lacks a direct equivalent topic that captures this specific domain. This highlights BERTopic's "COFFEE"-driven ability to detect gender-specific contributions in highly specialized and policy-relevant areas, reflecting the prominent role of women in public health professions, nursing, and health communication (Canadian Nurses Association, 2023).

Table 5.11: BERTopic Gender Effects: Public Health and Vaccine Communication

Topic	**Gender**	**Estimate**	**Std. Error**	**t-value**	**p-value**
Public Health & Vaccine Communication	Intercept	0.0186	0.0005	35.2883	<0.0001
	Female	0.0029	0.0006	5.0659	<0.0001

Overall, these findings illustrate COFFEE-powered BERTopic's nuanced capability to reveal gender-based research trends, offering a deeper understanding of how gender influences thematic contributions in various fields. By highlighting both disparities and areas where women are making significant impacts, this analysis can provide valuable insights that can inform policy decisions and initiatives aimed at promoting gender equity in research.

To examine the potential effects of gender on research funding, we conducted a comprehensive analysis on two sample topics: “Environmental Science & Industrial Processes” and “Quantum & Nuclear Physics”. Figure 5.4 demonstrates a gender disparity in the number of approved research proposals within the “Environmental Science & Industrial Processes” field. However, the median chart indicates that the awarded amounts for male and female researchers follow a relatively similar pattern over time.

Figure 5.4 also clearly shows that the number of approved applications received by male researchers has consistently been higher than that of female applicants, resulting in proportionally larger total award amounts. However, it is important to note that we do not have access to submission records, and the dataset only contains approved proposals. Therefore, we cannot compare the acceptance rates for male and female applicants.

Figure 5.5 illustrates the number of applicants and the median award amounts in the field of “Quantum & Nuclear Physics”. Unlike the previous topic, this figure reveals a greater disparity in median award amounts and a less consistent number of male applicants over time. Regardless of the two sample topics analyzed in this section, these findings highlight how automatically generated topics can uncover hidden disparities and patterns that might be overlooked in aggregate analyses. This approach can help guide targeted interventions to address concerns and inequities within specific research domains.

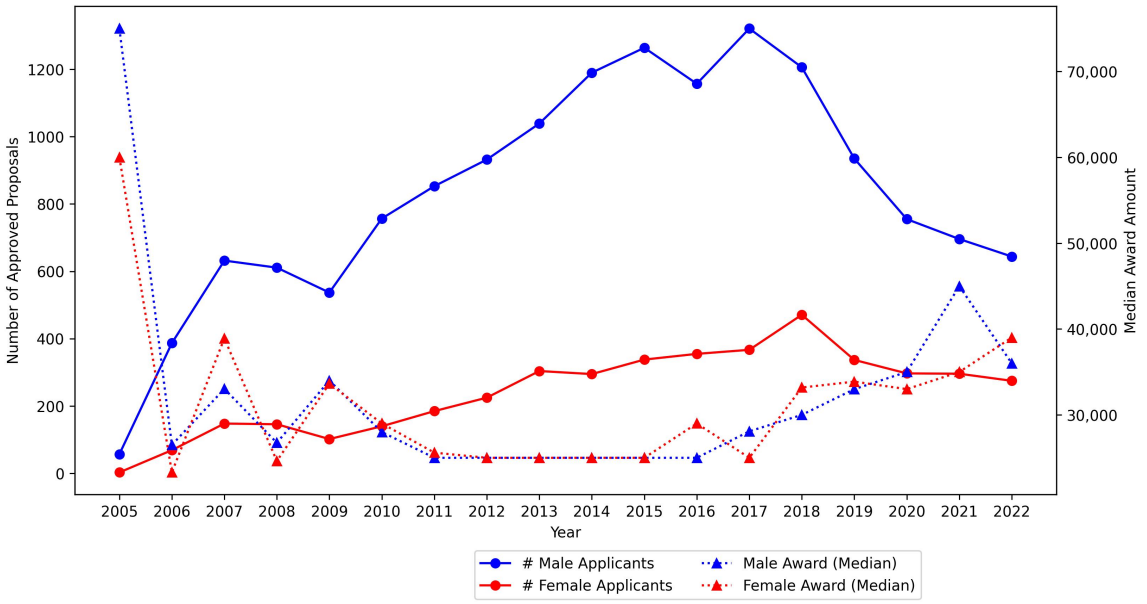

Figure 5.4: Trends in applicant numbers and median award amounts by gender for the "Environmental Science & Industrial Processes".

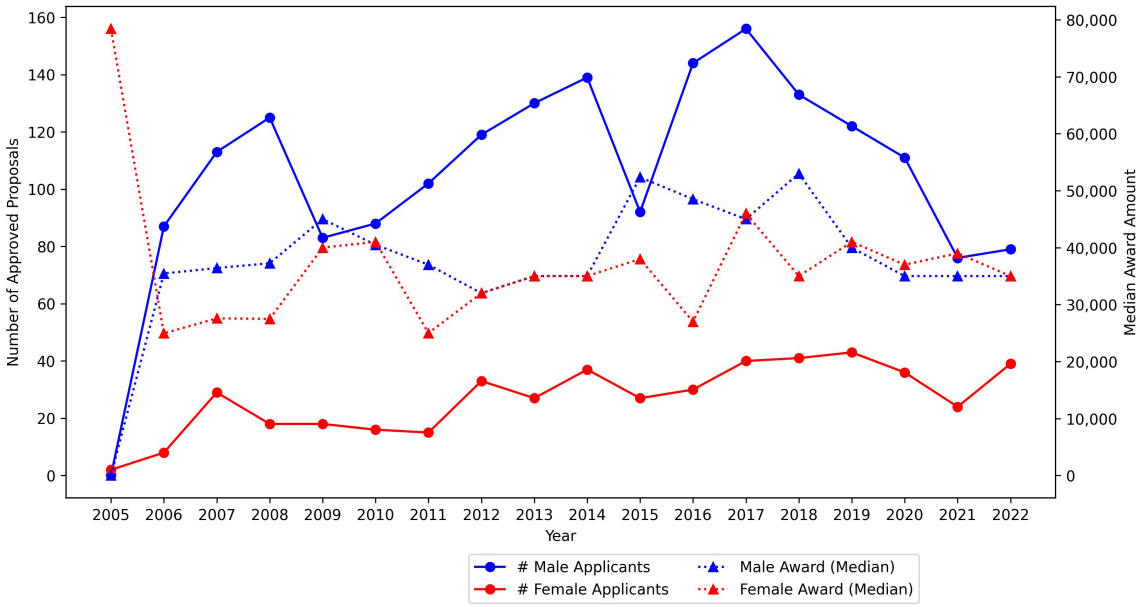

Figure 5.5: Trends in applicant numbers and median award amounts by gender for the "Quantum & Nuclear Physics".

Chapter 6

Discussion and Conclusion

This study presents a comprehensive comparative analysis of BERTopic, STM, and LDA, highlighting their distinct capabilities in uncovering thematic structures from a large corpus of Canadian NSERC-funded research proposals from 2005 to 2023. Our findings demonstrate that while all models robustly identify prominent scientific domains, their thematic resolution capabilities reveal distinct characteristics crucial for methodological selection.

BERTopic consistently demonstrated a superior ability to break down broad subjects into multiple, highly specialized, and semantically rich topics. This enhanced granularity allows it to provide a highly detailed and nuanced understanding of a domain by resolving distinct facets that might be blended into broader categories by other models. The contextual nature of its embeddings allows it to grasp subtle semantic relationships, leading to more insightful and coherent topic definitions, thus providing uniquely insightful and nuanced semantic discovery.

STM generally provided a good balance between identifying broad, well-established themes and reasonably distinct sub-topics. Its unique topics are coherent and well-defined, indicating a solid capability to identify meaningful clusters. It serves as a robust option when a moderate level of thematic granularity is desired.

Conversely, LDA, while effective at identifying main thematic currents and demonstrating good topic diversity for these broader themes, consistently produced more generalized topics. Its inherent probabilistic nature, relying on word co-occurrence, often results in topics that are less specific or

insightful for dissecting niche research areas, making it comparatively less effective for very fine-grained thematic analysis.

Collectively, this research offers methodological guidance for selecting appropriate topic modeling approaches based on desired analytical granularity, affirming that while all models robustly identify prominent scientific domains, BERTopic excels in uncovering more granular, highly specific, and often novel themes.

A central contribution of this work is the development and application of the COFFEE algorithm. This novel, bootstrap-based pipeline was designed to overcome the inferential limitations of non-probabilistic topic models, enabling robust statistical analysis of covariate effects. By pairing COFFEE with BERTopic and comparing the results to STM's established `estimateEffect` function and real-world, we first validated our approach by corroborating known research specializations. Both frameworks, for instance, identified Alberta's leadership in "Environmental Science & Industrial Processes" and the prominence of "Computer Science & Artificial Intelligence" in Ontario and British Columbia.

More importantly, this dual-model framework highlighted the superior analytical resolution of the BERTopic-COFFEE approach. It uncovered unique, statistically significant regional niches that were invisible to STM. For example, BERTopic uniquely detected a significant research focus on "Environmental Science" in Newfoundland and Labrador and on "Materials Science & Applied Physics" in New Brunswick. The analysis of gender effects was equally revealing. While both models confirmed the well-documented underrepresentation of women in "Computer Science & AI," BERTopic's granular topic resolution uniquely identified "Public Health & Vaccine Communication" as a field with a significant positive association with female researchers. This finding is particularly potent as STM could not even test this relationship, having failed to identify the topic in the first place.

These findings have significant implications for science policy. By providing a more granular and sensitive analytical tool, the BERTopic-COFFEE framework allows funding agencies like NSERC to move beyond high-level summaries toward a more nuanced understanding of the research landscape. Such precision is vital for developing targeted, evidence-based strategies that support regional research ecosystems and promote the goals of Equity, Diversity, and Inclusion.

Chapter 7

Limitations and Future Work

While this study provides valuable insights into funded research trends and the performance of various topic models, several limitations should be acknowledged. First, the quality of topic modelling outputs is significantly influenced by the preprocessing step involved. Although we ensured consistency across all models for direct comparison, variations in tokenization, lemmatization, and stop-word removal could affect the result. Future studies could explore the impact of different preprocessing techniques to assess their influence on model outcomes.

Second, the choice of the 0.82 cosine similarity threshold for defining topic similarity between models is a hyperparameter. This threshold was determined through an iterative qualitative assessment to optimize meaningful thematic correspondences. However, selecting a different threshold could alter the categorization of topics. Further research could examine the effects of varying this parameter and develop methods for dynamically adjusting it based on dataset characteristics.

Third, the interpretation and labelling of topics, initially generated using the GPT-4o model and refined through human review, are inherently subjective. Despite efforts to mitigate bias, different human interpretations and potential biases in large language models should be considered. Developing more objective and automated methods for topic interpretation could enhance reliability.

Methodologically, further investigation into robust, model-integrated covariate effect estimation techniques for embedding-based topic models is warranted to reduce reliance on post-hoc bootstrapping. Such advancements would improve the precision and reliability of effect estimation in embedding-based models, e.g., BERTopic.

Substantively, future work could expand the dataset to include research proposals funded by other organizations, such as Canadian Tri-Agencies (Canadian Institutes of Health Research(CIHR), Social Sciences and Humanities Research Council of Canada (SSHRC)). This would provide a more comprehensive view of the national research ecosystem. Additionally, a deeper exploration of the underlying factors driving provincial and gender-based specializations, potentially integrating additional demographic variables or historical policy analyses, would offer valuable context.

Finally, exploring the temporal evolution of these patterns in greater detail and investigating the relationship between topic prevalence and actual funding outcomes (e.g., success rates, award amounts) could yield critical insights for optimizing science policy and fostering a more inclusive and innovative scientific community.

Appendix A

A.1 Detailed Regression Results

Table A.1: BERTopic Provincial Effects: Regression Coefficients (Estimate, Std. Error, t-value, and p-value)

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
Environmental Science & Industrial Processes	Intercept	0.0568	0.0014	39.6711	<0001
	Alberta	0.0229	0.0028	8.1083	<0001
	British Columbia	0.0062	0.0022	2.8072	0.0050
	Manitoba	−0.0003	0.0038	−0.0818	0.9348
	New Brunswick	−0.0027	0.0057	−0.4749	0.6348
	Newfoundland and Labrador	0.0128	0.0056	2.2831	0.0224
	Nova Scotia	0.0006	0.0039	0.1566	0.8756
	Ontario	−0.0042	0.0021	−2.0046	0.0450
	Other	−0.0477	0.0021	−22.7762	<0001
	Quebec	−0.0002	0.0021	−0.0934	0.9255
Computer Science & Artificial Intelligence	Intercept	0.0449	0.0015	30.8181	<0001
	Alberta	0.0003	0.0024	0.1369	0.8911

Table A.1: Continued from previous page

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
	British Columbia	0.0051	0.0023	2.2319	0.0256
	Manitoba	0.0011	0.0037	0.3078	0.7583
	New Brunswick	0.0045	0.0066	0.6816	0.4955
	Newfoundland and Labrador	0.0017	0.0051	0.3329	0.7392
	Nova Scotia	−0.0103	0.0034	−3.0367	0.0024
	Ontario	0.0295	0.0018	5.1528	<0001
	Other	−0.0151	0.0030	−5.0668	<0001
	Quebec	−0.0046	0.0015	−3.1233	0.0018
Molecular Biology & Biotechnology	Intercept	0.0586	0.0010	56.2270	<0001
	Alberta	−0.0063	0.0018	−3.4805	0.0005
	British Columbia	−0.0019	0.0024	−0.7730	0.4395
	Manitoba	0.0206	0.0042	4.8793	<0001
	New Brunswick	0.0140	0.0062	2.2621	0.0237
	Newfoundland and Labrador	−0.0056	0.0048	−1.1761	0.2396
	Nova Scotia	0.0069	0.0031	2.2466	0.0247
	Ontario	−0.0083	0.0015	−5.3439	<0001
	Other	−0.0343	0.0034	−10.0927	<0001
	Quebec	−0.0097	0.0015	−6.2402	<0001
Materials Science & Applied Physics (Imaging & Photonics)	Intercept	0.0292	0.0007	40.5076	<0001
	Alberta	−0.0007	0.0017	−0.3927	0.6945

Table A.1: Continued from previous page

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
	British Columbia	0.0015	0.0015	0.9940	0.3202
	Manitoba	0.0042	0.0040	1.0559	0.2910
	New Brunswick	0.0181	0.0035	5.1810	<0001
	Newfoundland and Labrador	0.0017	0.0042	0.3983	0.6904
	Nova Scotia	−0.0005	0.0026	−0.1872	0.8515
	Ontario	0.0008	0.0012	0.6583	0.5103
	Other	−0.0208	0.0015	−13.7424	<0001
	Quebec	−0.0009	0.0013	−0.7276	0.4668
Neuroscience & Cognitive Science	Intercept	0.0405	0.0010	40.3336	<0001
	Alberta	0.0061	0.0024	2.5869	0.0097
	British Columbia	0.0021	0.0020	1.0530	0.2924
	Manitoba	0.0086	0.0039	2.2157	0.0267
	New Brunswick	−0.0127	0.0042	−3.0541	0.0023
	Newfoundland and Labrador	0.0165	0.0063	2.6373	0.0084
	Nova Scotia	0.0039	0.0038	1.0240	0.3058
	Ontario	0.0124	0.0016	7.9094	<0001
	Other	−0.0233	0.0025	−9.2669	<0001
	Quebec	0.0038	0.0013	2.8560	0.0043
Quantum & Nuclear Physics	Intercept	0.0544	0.0015	36.6840	<0001
	Alberta	−0.0063	0.0030	−2.0952	0.0362
	British Columbia	0.0125	0.0025	4.9586	<0001

Table A.1: Continued from previous page

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
	Manitoba	−0.0006	0.0041	−0.1368	0.8912
	New Brunswick	0.0059	0.0063	0.9344	0.3501
	Newfoundland and Labrador	−0.0113	0.0051	−2.1993	0.0279
	Nova Scotia	−0.0093	0.0028	−3.2934	0.0010
	Ontario	0.0056	0.0018	3.1644	0.0016
	Other	0.0358	0.0045	7.9018	<0001
	Quebec	−0.0153	0.0020	−7.5679	<0001
Musculoskeletal Biomechanics	Intercept	0.0202	0.0009	23.2768	<0001
	Alberta	0.0024	0.0015	1.6532	0.0983
	British Columbia	0.0018	0.0016	1.1002	0.2712
	Manitoba	0.0044	0.0024	1.8389	0.0659
	New Brunswick	0.0055	0.0029	1.8906	0.0587
	Newfoundland and Labrador	0.0003	0.0035	0.0983	0.9217
	Nova Scotia	−0.0019	0.0023	−0.8302	0.4064
	Ontario	0.0059	0.0010	5.7589	<0001
	Other	−0.0157	0.0010	−15.6707	<0001
	Quebec	0.0012	0.0011	1.1599	0.2461
Dental & Oral Biomaterials	Intercept	0.0074	0.0005	16.3041	<0001
	Alberta	−0.0003	0.0008	−0.3287	0.7424
	British Columbia	−0.0004	0.0006	−0.6092	0.5424
	Manitoba	0.0011	0.0014	0.8240	0.4099

Table A.1: Continued from previous page

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
	New Brunswick	−0.0039	0.0013	−2.9155	0.0036
	Newfoundland and Labrador	−0.0037	0.0014	−2.5552	0.0106
	Nova Scotia	0.0046	0.0018	2.5655	0.0103
	Ontario	0.0019	0.0007	2.7998	0.0051
	Other	−0.0046	0.0010	−4.8555	<0001
	Quebec	0.0040	0.0008	5.2781	<0001
Public Health & Vaccine Communication	Intercept	0.0154	0.0007	22.1753	<0001
	Alberta	0.0042	0.0014	3.0594	0.0022
	British Columbia	0.0015	0.0014	1.0831	0.2788
	Manitoba	0.0018	0.0022	0.7915	0.4286
	New Brunswick	−0.0053	0.0024	−2.1951	0.0282
	Newfoundland and Labrador	−0.0125	0.0013	−9.4260	<0001
	Nova Scotia	0.0026	0.0018	1.4691	0.1418
	Ontario	0.0047	0.0010	4.6628	<0001
	Other	−0.0061	0.0022	−2.7444	0.0061
	Quebec	−0.0021	0.0010	−2.1242	0.0337
Mechanical Engineering & Sports Technology	Intercept	0.0043	0.0003	13.5216	<0001
	Alberta	−0.0001	0.0008	−0.1836	0.8544
	British Columbia	0.0016	0.0007	2.2140	0.0268
	Manitoba	0.0013	0.0016	0.8185	0.4131
	New Brunswick	0.0033	0.0023	1.4372	0.1507

Table A.1: Continued from previous page

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
	Newfoundland and Labrador	−0.0037	0.0005	−7.3070	<0001
	Nova Scotia	−0.0023	0.0008	−2.9480	0.0032
	Ontario	0.0002	0.0004	0.3815	0.7028
	Other	−0.0043	0.0003	−13.5216	<0001
	Quebec	0.0060	0.0006	9.6339	<0001
Food Science & Agricultural Products	Intercept	0.0607	0.0015	39.4140	<0001
	Alberta	0.0341	0.0027	12.4820	<0001
	British Columbia	−0.0302	0.0022	−13.5390	<0001
	Manitoba	0.0200	0.0043	4.6310	<0001
	New Brunswick	−0.0133	0.0049	−2.6965	0.0070
	Newfoundland and Labrador	0.0344	0.0073	4.7343	<0001
	Nova Scotia	−0.0010	0.0047	−0.2208	0.8252
	Ontario	−0.0248	0.0017	−14.2935	<0001
	Other	−0.0311	0.0030	−10.2104	<0001
	Quebec	−0.0189	0.0022	−8.7359	<0001
Advanced Energy Storage & Electronics	Intercept	0.0341	0.0010	33.1378	<0001
	Alberta	0.0001	0.0015	0.0563	0.9551
	British Columbia	0.0034	0.0017	1.9710	0.0487
	Manitoba	−0.0062	0.0034	−1.8373	0.0662
	New Brunswick	−0.0076	0.0047	−1.6201	0.1052
	Newfoundland and Labrador	−0.0009	0.0046	−0.2022	0.8397

Table A.1: Continued from previous page

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
	Nova Scotia	0.0130	0.0043	2.9984	0.0027
	Ontario	0.0103	0.0012	8.8871	<0001
	Other	−0.0088	0.0028	−3.0926	0.0020
	Quebec	0.0040	0.0012	3.2651	0.0011
Organ Transplantation & Biofabrication	Intercept	0.0162	0.0008	21.4995	<0001
	Alberta	0.0011	0.0015	0.7286	0.4663
	British Columbia	0.0015	0.0008	1.8710	0.0614
	Manitoba	0.0042	0.0027	1.5529	0.1204
	New Brunswick	−0.0040	0.0030	−1.3184	0.1874
	Newfoundland and Labrador	−0.0005	0.0033	−0.1655	0.8686
	Nova Scotia	0.0002	0.0020	0.1174	0.9065
	Ontario	0.0050	0.0010	4.9844	<0001
	Other	−0.0081	0.0017	−4.7325	<0001
	Quebec	−0.0002	0.0010	−0.1938	0.8463

Note: p-values < 0.0001 are reported as such due to precision limits.

Table A.2: STM Provincial Effects: Regression Coefficients (Estimate, Std. Error, t-value, and p-value)

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
Molecular Biology & Biotechnology	Intercept	0.0854	0.0013	65.8220	<0001
	Alberta	0.0057	0.0026	2.2107	0.0271
	British Columbia	−0.0055	0.0023	−2.3882	0.0169
	Manitoba	0.0230	0.0047	4.9114	<0001
	New Brunswick	−0.0331	0.0055	−5.9755	<0001
	Newfoundland and Labrador	−0.0085	0.0057	−1.4861	0.1373
	Nova Scotia	−0.0001	0.0038	−0.0241	0.9807
	Ontario	0.0040	0.0017	2.3675	0.0179
	Other	0.0068	0.0046	1.4731	0.1407
	Québec	−0.0001	0.0020	−0.0676	0.9461
Climate & Aquatic Ecology	Intercept	0.0955	0.0012	81.4957	<0001
	Alberta	−0.0106	0.0021	−4.9708	<0001
	British Columbia	−0.0029	0.0022	−1.3255	0.1850
	Manitoba	0.0009	0.0037	0.2366	0.8130
	New Brunswick	0.0086	0.0052	1.6676	0.0954
	Newfoundland and Labrador	0.0402	0.0059	6.8622	<0001
	Nova Scotia	0.0194	0.0039	5.0056	<0001
	Ontario	−0.0243	0.0015	−16.0212	<0001
	Other	−0.0244	0.0040	−6.1851	<0001
	Québec	−0.0162	0.0016	−10.3202	<0001

Table A.2: Continued from previous page

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
Imaging & Sensing Applications	Intercept	0.0820	0.0009	90.1221	<0001
	Alberta	0.0053	0.0018	2.9428	0.0033
	British Columbia	0.0113	0.0016	6.8706	<0001
	Manitoba	0.0043	0.0031	1.3849	0.1661
	New Brunswick	−0.0042	0.0039	−1.0796	0.2803
	Newfoundland and Labrador	−0.0104	0.0043	−2.4385	0.0148
	Nova Scotia	−0.0001	0.0028	−0.0236	0.9812
	Ontario	0.0123	0.0012	10.1497	<0001
	Other	−0.0099	0.0031	−3.1750	0.0015
	Québec	0.0024	0.0013	1.7872	0.0739
Energy & Advanced Materials	Intercept	0.0341	0.0010	33.1378	<0001
	Alberta	0.0001	0.0015	0.0563	0.9551
	British Columbia	0.0034	0.0017	1.9710	0.0487
	Manitoba	−0.0062	0.0034	−1.8373	0.0662
	New Brunswick	−0.0076	0.0047	−1.6201	0.1052
	Newfoundland and Labrador	−0.0009	0.0046	−0.2022	0.8397
	Nova Scotia	0.0130	0.0043	2.9984	0.0027
	Ontario	0.0103	0.0012	8.8871	<0001
	Other	−0.0088	0.0028	−3.0926	0.0020
	Québec	0.0040	0.0012	3.2651	0.0011

Table A.2: Continued from previous page

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
Population & Evolutionary Biology	Intercept	0.0702	0.0010	70.5993	<0001
	Alberta	−0.0113	0.0018	−6.1731	<0001
	British Columbia	−0.0059	0.0016	−3.6148	0.0003
	Manitoba	−0.0053	0.0031	−1.7059	0.0880
	New Brunswick	0.0180	0.0043	4.1779	<0001
	Newfoundland and Labrador	0.0164	0.0048	3.4216	0.0006
	Nova Scotia	0.0100	0.0029	3.4689	0.0005
	Ontario	−0.0147	0.0013	−11.3717	<0001
	Other	0.0086	0.0035	2.4704	0.0135
	Québec	−0.0196	0.0014	−13.8604	<0001
Plant Pathology & Crop Genetics	Intercept	0.0604	0.0009	65.0561	<0001
	Alberta	−0.0042	0.0018	−2.3327	0.0197
	British Columbia	−0.0079	0.0015	−5.2530	<0001
	Manitoba	0.0153	0.0033	4.6657	<0001
	New Brunswick	−0.0076	0.0040	−1.8763	0.0606
	Newfoundland and Labrador	−0.0095	0.0041	−2.3193	0.0204
	Nova Scotia	0.0040	0.0029	1.3912	0.1642
	Ontario	−0.0063	0.0012	−5.3607	<0001
	Other	−0.0064	0.0032	−2.0065	0.0448
	Québec	−0.0034	0.0013	−2.6103	0.0090

Table A.2: Continued from previous page

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
Environmental Science & Industrial Processes	Intercept	0.0998	0.0012	81.6671	<0001
	Alberta	0.0202	0.0024	8.5651	<0001
	British Columbia	−0.0195	0.0020	−9.8015	<0001
	Manitoba	0.0000	0.0042	0.0032	0.9974
	New Brunswick	−0.0081	0.0054	−1.5000	0.1336
	Newfoundland and Labrador	0.0070	0.0054	1.2952	0.1952
	Nova Scotia	0.0013	0.0037	0.3453	0.7298
	Ontario	−0.0166	0.0015	−10.7613	<0001
	Other	−0.0260	0.0040	−6.4873	<0001
	Québec	0.0183	0.0018	10.1931	<0001
Computer Science & Artificial Intelligence	Intercept	0.1171	0.0016	74.8581	<0001
	Alberta	0.0023	0.0028	0.8288	0.4072
	British Columbia	0.0120	0.0025	4.7339	<0001
	Manitoba	−0.0109	0.0048	−2.2566	0.0240
	New Brunswick	0.0091	0.0063	1.4561	0.1454
	Newfoundland and Labrador	−0.0115	0.0068	−1.6862	0.0918
	Nova Scotia	−0.0074	0.0042	−1.7465	0.0807
	Ontario	0.0239	0.0019	12.8574	<0001
	Other	−0.0260	0.0048	−5.4103	<0001
	Québec	0.0159	0.0021	7.4986	<0001

Table A.2: Continued from previous page

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
Neuroscience & Cognitive Science	Intercept	0.0756	0.0012	60.8447	<0001
	Alberta	−0.0018	0.0022	−0.7986	0.4245
	British Columbia	0.0037	0.0021	1.7847	0.0743
	Manitoba	−0.0097	0.0039	−2.5146	0.0119
	New Brunswick	−0.0039	0.0050	−0.7806	0.4350
	Newfoundland and Labrador	0.0023	0.0054	0.4260	0.6701
	Nova Scotia	−0.0009	0.0035	−0.2509	0.8019
	Ontario	0.0106	0.0016	6.8261	<0001
	Other	0.0157	0.0040	3.8920	<0001
	Québec	−0.0024	0.0016	−1.4524	0.1464
Materials Science & Applied Physics (Imaging & Photonics)	Intercept	0.0876	0.0012	71.2635	<0001
	Alberta	−0.0059	0.0022	−2.6963	0.0070
	British Columbia	0.0080	0.0020	3.9716	<0001
	Manitoba	0.0005	0.0042	0.1072	0.9147
	New Brunswick	−0.0021	0.0054	−0.3917	0.6952
	Newfoundland and Labrador	−0.0181	0.0053	−3.4106	0.0006
	Nova Scotia	−0.0070	0.0036	−1.9766	0.0481
	Ontario	−0.0009	0.0016	−0.5419	0.5879
	Other	0.0421	0.0040	10.5007	<0001
	Québec	−0.0069	0.0018	−3.9349	<0001

Table A.2: Continued from previous page

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
Theoretical & Computational Sciences	Intercept	0.1202	0.0012	96.8037	<0001
	Alberta	−0.0065	0.0023	−2.8836	0.0039
	British Columbia	0.0010	0.0021	0.4775	0.6330
	Manitoba	−0.0075	0.0042	−1.7682	0.0770
	New Brunswick	0.0136	0.0055	2.4861	0.0129
	Newfoundland and Labrador	0.0176	0.0055	3.1849	0.0014
	Nova Scotia	−0.0146	0.0036	−4.0220	<0001
	Ontario	0.0008	0.0016	0.5222	0.6015
	Other	0.0179	0.0042	4.2388	<0001
	Québec	−0.0052	0.0018	−2.8892	0.0039

Note: p-values < 0.0001 are reported as such due to precision limits.

Table A.3: BERTopic Gender Effects: Regression Coefficients (Estimate, Std. Error, t-value, and p-value)

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
Environmental Science & Industrial Processes	Intercept	0.0565	0.0008	71.4637	<0001
	Female	−0.0022	0.0009	−2.5313	0.0114
Computer Science & Artificial Intelligence	Intercept	0.0461	0.0008	58.0180	<0001
	Female	−0.0034	0.0008	−4.2460	<0001
Molecular Biology & Biotechnology	Intercept	0.0545	0.0009	61.6490	<0001
	Female	0.0033	0.0009	3.5601	0.0004
Materials Science & Applied Physics (Imaging & Photonics)	Intercept	0.0274	0.0005	55.4532	<0001
	Female	−0.0037	0.0005	−6.9179	<0001
Neuroscience & Cognitive Science	Intercept	0.0508	0.0009	55.8000	<0001
	Female	0.0086	0.0008	10.7550	<0001
Quantum & Nuclear Physics	Intercept	0.0486	0.0007	70.1653	<0001
	Female	−0.0104	0.0008	−13.8659	<0001
Musculoskeletal Biomechanics	Intercept	0.0234	0.0006	36.7802	<0001
	Female	0.0013	0.0005	2.7189	0.0066
Dental & Oral Biomaterials	Intercept	0.0092	0.0003	27.2377	<0001
	Female	0.0003	0.0003	0.7716	0.4403

Table A.3: Continued from previous page

BERTopic Topic Name	Province	Estimate	Std. Error	t-value	p-value
Public Health & Vaccine Communication	Intercept	0.0186	0.0005	35.2883	<0001
	Female	0.0029	0.0006	5.0659	<0001
Mechanical Engineering & Sports Technology	Intercept	0.0055	0.0003	16.8807	<0001
	Female	−0.0006	0.0003	−1.6514	0.0987
Food Science & Agricultural Products	Intercept	0.0476	0.0009	54.1775	<0001
	Female	0.0021	0.0008	2.5745	0.0100
Advanced Energy Storage & Electronics	Intercept	0.0363	0.0006	59.0322	<0001
	Female	−0.0065	0.0006	−10.3217	<0001
Organ Transplantation & Biofabrication	Intercept	0.0192	0.0005	37.9411	<0001
	Female	0.0019	0.0005	3.7516	0.0002

Note: p-values < 0.0001 are reported as such due to precision limits.

Table A.4: STM Gender Effects: Regression Coefficients (Estimate, Std. Error, t-value, and p-value)

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
Molecular Biology & Biotechnology	Intercept	0.1019	0.0015	66.5039	<0001
	Female	0.0200	0.0017	11.4811	<0001
Climate & Aquatic Ecology	Intercept	0.0927	0.0013	70.4617	<0001
	Female	0.0148	0.0015	9.7751	<0001
Imaging & Sensing Applications	Intercept	0.0842	0.0011	79.2597	<0001
	Female	−0.0061	0.0013	−4.7253	<0001
Energy & Advanced Materials	Intercept	0.0999	0.0015	67.9884	<0001
	Female	−0.0209	0.0017	−12.4113	<0001
Population & Evolutionary Biology	Intercept	0.0730	0.0011	64.5050	<0001
	Female	0.0192	0.0013	14.3621	<0001
Plant Pathology & Crop Genetics	Intercept	0.0643	0.0010	63.3346	<0001
	Female	0.0103	0.0012	8.8583	<0001
Environmental Science & Industrial Processes	Intercept	0.1033	0.0014	72.4233	<0001
	Female	0.0074	0.0017	4.3828	<0001
Computer Science & Artificial Intelligence	Intercept	0.1186	0.0015	78.8863	<0001
	Female	−0.0165	0.0018	−9.2835	<0001
Neuroscience & Cognitive Science	Intercept	0.0917	0.0013	68.0851	<0001

Table A.4: Continued from previous page

STM Topic Name	Province	Estimate	Std. Error	t-value	p-value
	Female	0.0170	0.0016	10.6382	<0001
Materials Science & Applied Physics (Imaging & Photonics)	Intercept	0.0706	0.0013	54.3399	<0001
	Female	−0.0207	0.0015	−13.6727	<0001
Theoretical & Computational Sciences	Intercept	0.0999	0.0014	71.2831	<0001
	Female	−0.0246	0.0016	−15.4314	<0001

Note: p-values < 0.0001 are reported as such due to precision limits.

References

Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). Gender differences in research collaboration. *Journal of Informetrics*, *7*(4), 811–822.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2024). *Gpt-4 technical report.* Retrieved from https://arxiv.org/abs/2303.08774

Arnaout, A., Gill, P., Virani, A., Flatt, A., Prodan-Balla, N., Byres, D., ... Virani, S. (2024). Shaping the future of healthcare in british columbia: Establishing provincial clinical governance for responsible deployment of artificial intelligence tools. *Healthcare Management Forum*, *37*(5), 320–328. Retrieved from https://journals.sagepub.com/home/hmf doi: 10.1177/08404704241264819

Asheim, B., Grillitsch, M., & Trippl, M. (2016). Regional innovation systems: past - presence - future. In D. Doloreux, R. Shearmur, & C. Carrincazeaux (Eds.), *Handbook on the geographies of innovation* (pp. 45–62). Edward Elgar Publishing. doi: 10.4337/9781784710774.00010

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (p. 113–120). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1143844.1143859 doi: 10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2007a). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226–238.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2007b). Gender differences in grant peer review: A

meta-analysis. *Journal of Informetrics*, *1*(3), 226–238.

Boschma, R. (2005). Proximity and innovation: a critical assessment. *Regional studies*, *39*(1), 61–74. doi: 10.1080/0034340052000320887

Breschi, S., Lenzi, C., Lissoni, F., & Vezzulli, A. (2010). 16 the geography of knowledge spillovers: the role of inventors' mobility across firms and in space. *The handbook of evolutionary economic geography*, 353. Retrieved from https://doi.org/10.1093/jeg/lbp049 doi: 10.1093/jeg/lbp049

Burgelman, J.-C., Chloupková, J., & Wobbe, W. (2014). Foresight in support of european research and innovation policies: The european commission is preparing the funding of grand societal challenges. *European Journal of Futures Research*, *2*(1), 55.

Burney, S., Donelle, L., & Kothari, A. (2025). Exploring the public health agency of canada's and the ontario government's vaccine-related crisis communication on x during the covid-19 pandemic. *FACETS*, *10*, 1–16. Retrieved from https://doi.org/10.1139/facets-2022-0186 doi: 10.1139/facets-2022-0186

Canadian Nurses Association. (2023). *Nursing statistics.* Retrieved from https://www.cna-aiic.ca/en/nursing/regulated-nursing-in-canada/nursing-statistics (Accessed: 2025-07-10)

Canadian Society for Molecular Biosciences. (2024). *2024 Annual Conference – Canadian Society for Molecular Biosciences.* https://www.csmb-scbm.ca/meetings/2024-annual-conference/. (Accessed)

Chen, X., Xie, H., Tao, X., Xu, L., Wang, J., Dai, H.-N., & Wang, F. L. (2024). A topic modeling-based bibliometric exploration of automatic summarization research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(5), e1540.

Deanna, R., Merkle, B. G., Chun, K. P., Navarro-Rosenblatt, D., Baxter, I., Oleas, N., ... others (2022). Community voices: the importance of diverse networks in academic mentoring. *Nature Communications*, *13*(1), 1681.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423/ doi: 10.18653/v1/N19-1423

Dubé, M. G., Dunlop, J. M., Davidson, C., Beausoleil, D. L., Hazewinkel, R. R., & Wyatt, F. (2021). History, overview, and governance of environmental monitoring in the oil sands region of alberta, canada. *Integrated Environmental Assessment and Management*, *18*(2), 319–332. doi: 10.1002/ieam.4490

Ebadi, A., Tremblay, S., Goutte, C., & Schiffauerova, A. (2020). Application of machine learning techniques to assess the trends and alignment of the funded research output. *Journal of Informetrics*, *14*(2), 101018.

Ebadi, A., Zahedi, M. R., Jowkar, M., & Zare, A. (2016). How to boost scientific production? a statistical analysis of research funding and other influencing factors. *Scientometrics*, *106*(3), 1117–1135.

Ecklund, E. H., Lincoln, A. E., & Tansey, C. (2012). Gender segregation in elite academic science. *Gender & Society*, *26*(5), 693–717.

Efron, B., & Narasimhan, B. (2020). The automatic construction of bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, *29*(3), 608–619. Retrieved from https://doi.org/10.1080/10618600.2020.1714633 (PMID: 33727780) doi: 10.1080/10618600.2020.1714633

Egger, R., & Yu, J. (2022, may 6). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, *7*, 886498. doi: 10.3389/fsoc.2022.886498

Florida, R. (2002). The economic geography of talent. *Annals of the Association of American geographers*, *92*(4), 743–755. doi: 10.1111/1467-8306.00325

Glenny, V., Tuke, J., Bean, N., & Mitchell, L. (2019). *A framework for streamlined statistical prediction using topic models.* Retrieved from https://arxiv.org/abs/1904.06941

Gray, T. S. (Ed.). (2005). *Participation in fisheries governance* (Vol. 4). Dordrecht, The Netherlands: Springer. (Jennifer L. Nielsen is the Series Editor [1])

Griffiths, T., Jordan, M., Tenenbaum, J., & Blei, D. (2003). Hierarchical topic models and the nested chinese restaurant process. In (Vol. 16).

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Grootendorst, M. (2020). Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. *Zenodo, Version v0*, *9*(10.5281).

Hajibabaei, A., Schiffauerova, A., & Ebadi, A. (2022). Gender-specific patterns in the artificial intelligence scientific ecosystem. *Journal of Informetrics*, *16*, 101275.

Hajibabaei, A., Schiffauerova, A., & Ebadi, A. (2023). Women and key positions in scientific collaboration networks: analyzing central scientists' profiles in the artificial intelligence ecosystem through a gender lens. *Scientometrics*, *128*, 1219-1240. doi: 10.1007/s11192-022-04601-5

Hango, D. (2013, december). Gender differences in science, technology, engineering, mathematics and computer science (stem) programs at university. , 1-11. Retrieved from https://www150.statcan.gc.ca/n1/pub/75-006-x/2013001/article/11874-eng.pdf

Hankar, M., Kasri, M., & Beni-Hssane, A. (2025). A comprehensive overview of topic modeling: Techniques, applications and challenges. *Neurocomputing*, 129638.

Healy, J., & McInnes, L. (2024). Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, *4*(1), 82.

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, *95*(9-10), 1168–1177.

Jaramillo, A. M., Macedo, M., Oliveira, M., Karimi, F., & Menezes, R. (2025). *Systematic comparison of gender inequality in scientific rankings across disciplines.* Retrieved from https://arxiv.org/abs/2501.13061

Lang, R., Benham, J. L., Atabati, O., & et al. (2021). Attitudes, behaviours and barriers to public health measures for covid-19: a survey to inform public health messaging. *BMC Public Health*, *21*(1), 765. Retrieved from https://doi.org/10.1186/s12889-021-10790-0 doi: 10.1186/s12889-021-10790-0

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Macnaghten, P. (2022). Models of science policy: from the linear model to responsible research and innovation. In *The responsibility of science* (pp. 93–106). Springer International Publishing Cham.

McCann, P. (2001). *Urban and regional economics*. Oxford University Press.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205. doi: 10.21105/joss.00205

Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. Retrieved from https://www.science.org/doi/abs/10.1126/science.159.3810.56 doi: 10.1126/science.159.3810.56

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In R. Barzilay & M. Johnson (Eds.), *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D11-1024/

Mooney, C. Z. (1996). Bootstrap statistical inference: Examples and evaluations for political science. *American Journal of Political Science*, *40*(2), 570–602.

Natural Resources Canada. (2024). *About canmetmaterials.* Natural Resources Canada website, Government of Canada. Retrieved from https://natural-resources.canada.ca/science-data/science-research/research-centres/canmetmaterials (Date Modified: 2024-12-20)

Natural Sciences and Engineering Research Council of Canada. (2022, September 21). *Nserc 2030: Discovery. innovation. inclusion.* Online. Retrieved from https://www.nserc-crsng.gc.ca/nserc-crsng/nserc2030-crsng2030/report-rapport/index_eng.asp (Accessed 2025-07-20)

Ogden, L. E. (2019, nov). Study finds gender differences in success rates for canadian scientific research grants. *University Affairs*. Retrieved from https://universityaffairs.ca/news/study-finds-gender-differences-in-success-rates-for-canadian-scientific-research-grants/ (Accessed: October 24, 2025)

Omenn, G. S. (2006). Grand challenges and great opportunities in science, technology, and public policy. *Science*, *314*(5806), 1696–1704.

Petersen, O. H. (2021). Inequality of research funding between different countries and regions is a serious problem for global science. *Function (Oxf.)*, *2*(6), zqab060. doi: 10.1093/function/zqab060

Porter, M. E. (2008). *On competition*. Harvard Business Press.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An r package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. doi: 10.18637/jss.v091.i02

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. In *Proceedings of the nips 2013 workshop on topic models: Computation, application, and evaluation.* Retrieved from https://scholar.harvard.edu/files/dtingley/files/stmnips2013.pdf (Prepared for the NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation)

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (p. 399–408). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2684822.2685324 doi: 10.1145/2684822.2685324

Rodríguez-Pose, A. (2018). The revenge of the places that don't matter (and what to do about it). *Cambridge Journal of Regions, Economy and Society*, *11*(1), 189–209.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). The author-topic model for authors and documents..

Sato, S., Gygax, P. M., Randall, J., & Schmid Mast, M. (2021). The leaky pipeline in research grant peer review and funding decisions: challenges and future directions. *Higher Education*, *82*(1), 145–162. Retrieved from https://doi.org/10.1007/s10734-020-00626-y doi: 10.1007/s10734-020-00626-y

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants. *Sex Roles*, *57*(7), 509–514. Retrieved from https://doi.org/10.1007/s11199-007-9291-4 doi: 10.1007/s11199-007-9291-4

Schmaling, K. B., & Gallo, S. A. (2023). Gender differences in peer reviewed grant applications, awards, and amounts: a systematic review and meta-analysis. *Research integrity and peer review*, *8*(1), 2.

Schulze, P., Wiegrebe, S., Thurner, P. W., Heumann, C., & Aßenmacher, M. (2024). *A bayesian approach to modeling topic-metadata relationships* (Vol. 108) (No. 2). Springer.

Smith, R. D., Schäfer, S., & Bernstein, M. J. (2024). Governing beyond the project: Refocusing innovation governance in emerging science and technology funding. *Social Studies of Science*, *54*(3), 377–404.

Stephan, P. E. (2015). *How economics shapes science*. Harvard University Press.

Tibshirani, R. J., & Efron, B. (1993). *An introduction to the bootstrap* (Vol. 57) (No. 1). Retrieved from https://www.taylorfrancis.com/books/9781000064988

Torres, I. L., Collins, R.-N., Hertz, A., & Liukkonen, M. (2024). Policy proposals to promote inclusion of caregivers in the research funding system. *Frontiers in Education*, *Volume 9 - 2024*. Retrieved from https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1472517 doi: 10.3389/feduc.2024.1472517

Tri-agency equity, diversity, and inclusion action plan. (2021). Retrieved from https://www.nserc-crsng.gc.ca/NSERC-CRSNG/EDI-EDI/index_eng.asp

Van Arensbergen, P., Van der Weijden, I., & Van den Besselaar, P. (2012). Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, *93*(3), 857–868.

van den Besselaar, P., & Mom, C. (2022). Gender differences in research grant allocation–a mixed picture. *arXiv preprint arXiv:2205.13641*, *126*(4), 3191–3215.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Vector Institute for Artificial Intelligence. (2025). *Vector Institute for Artificial Intelligence.* https://vectorinstitute.ai/. (Accessed: October 25, 2025)

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer review. *Nature*, *387*(6631), 341–343.

Wilson, C. (2022). Public engagement and ai: A values analysis of national strategies. *Government Information Quarterly*, *39*(1), 101652. doi: 10.1016/j.giq.2024.101929

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? a natural experiment at a national funding agency. *The Lancet*, *393*(10171), 531–540.

Zhou, P., Cai, X., & Lyu, X. (2020). An in-depth analysis of government funding and international collaboration in scientific research. *Scientometrics*, *125*(2), 1331–1347.