Insights into India-Canada Relations: Analyzing Public Discourse on Facebook

By

Atodaria Vaidehi

A Major Research Project

presented to Toronto Metropolitan University

in partial fulfilment of the requirements for the degree of

Master of Science

in the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2024

Submitted To: **Dr. Anatoliy Gruzd, Dr. Ceni Babaoglu**

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I, Vaidehi Atodaria, am the sole author of this Major Research Paper (MRP). This submission truly represents my MRP, including any necessary final amendments.

I grant permission to Toronto Metropolitan University to lend this MRP to other institutions or individuals for scholarly research.

Additionally, I allow Toronto Metropolitan University to duplicate this MRP, partially or in full, through photocopying or other methods upon the request of other institutions or individuals for scholarly research.

I know my MRP may be accessible to the public in electronic format.

Vaidehi Atodaria

ACKNOWLEDGEMENTS

I am deeply thankful to Dr Anatoliy Gruzd for his exceptional guidance and support throughout my research project. As my supervisor, he has been an invaluable source of knowledge and inspiration. His expertise and insightful feedback have immensely contributed to my academic growth and the success of this project. I am genuinely grateful for the opportunity to learn from such a distinguished mentor. Thank you, **Dr. Gruzd,** for your dedication and invaluable contributions to my educational journey

I also express my sincere appreciation to **Dr. Babaoglu** for their insightful critiques and valuable suggestions, which greatly enhanced the quality of my work.

Thank you, Professor Anatoli and Dr.Ceni Babaoglu.

Table Of Contents

LIST OF FIGURES

1. Introduction

This study examines public discussions surrounding India-Canada relations on Facebook and their potential impact on individuals considering moving to Canada or those who have already settled here. With Facebook's expansive user base representing diverse perspectives and experiences, it is an ideal platform for this research. Drawing inspiration from a study by "Ngueleo et al. [1], which utilized natural language processing (NLP) to analyze Twitter posts on COVID-19 booster shots in India, this project employs comparable techniques. Specifically, it uses sentiment analysis and topic modelling to understand the sentiment and main topics shared by public Facebook accounts about India-Canada relations."

A. Background

The India-Canada relationship has seen its fair share of complexities, shaped by political, social, and trade factors. One of the significant issues causing tensions between the two countries is the Khalistan movement by some Sikhs, a religious minority in India, to create an independent state called "Khalistan in the Punjab region of India. This movement has a history of violence and is a sensitive subject in India [2]."

"This issue came to the forefront recently during the G20 Summit, held in New Delhi on September 9, [3]." India fears these activities could harm its sovereignty and security. Adding to the strain, a parade in "Brampton, Ontario, on June 4 featured a float that depicted the assassination of Indira Gandhi, a former Prime Minister of India, by her Sikh bodyguards. This event in Canada upset many people in India and made the situation more tense [4]." Because of these tensions, a trade mission from Canada to India was postponed in October, and talks about a trade deal were put on hold [5]."

B. Research Question

The central research question of the study is:

- To what extent does the public discourse on Facebook reflect the sentiments and critical topics surrounding the tensions between India and Canada?

- Can this analysis help us better understand public perceptions of immigration and the experiences of Indian expatriates in Canada?

The project leverages advanced computational methods, specifically natural language processing (NLP) techniques, to answer these questions. It aims to identify the sentiments expressed in these discussions—positive, negative, or neutral. Additionally, topic modelling categorizes the conversations into specific themes, revealing the primary debate topics.

2. Literature Review

Social media platforms, particularly Facebook, have emerged as vital spaces for public discourse. They offer insights into collective opinions and sentiments on various topics, including international relations and geopolitical issues. These platforms' capacity to reflect and shape public sentiment makes them invaluable for research, especially in understanding the dynamics of international disagreements as seen through the lens of everyday users.

In our study of social media discourse surrounding India-Canada issues, we selected TextBlob for sentiment analysis, drawing from the comprehensive insights provided in "A survey on sentiment analysis methods, applications, and challenges" by Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni [6]. This choice was informed by TextBlob's proven effectiveness in handling texts predominantly in English, aligning with the linguistic composition of our dataset, which includes English, Hindi, and Punjabi languages.

The simplicity and intuitive API of TextBlob enable efficient processing of large volumes of text data, which is crucial for our research needs. The literature emphasizes similar methodologies applied in multilingual social media data analysis, illustrating TextBlob's capability to accurately assess sentiment across various languages with minimal preprocessing, thus supporting our decision. This balance between computational efficiency and analytical depth, as supported by the referenced study, highlights TextBlob's suitability for our project, ensuring comprehensive sentiment analysis coverage across the diverse linguistic spectrum of our dataset.

Integrating TextBlob with NLP techniques, such as topic modelling, significantly enhances our analysis. Moreover, it provides detailed insights into social media conversations. This

multidimensional strategy was effectively employed in research conducted by A. Molenaar, D. Lukose, L. Brennan, E.L. Jenkins, and T.A. McCaffrey. Their study used NLP to examine "social media debates on food security by combining sentiment analysis with topic modelling to measure public sentiment and identify critical topics" [7]. This approach shows how combining these methodologies can provide in-depth insights across various subjects, giving detailed insights into social media conversations.

In our research into the dynamics of social media discussions concerning India-Canada relations, as observed on the social media platform Facebook, we purposely chose Latent Dirichlet Allocation (LDA) for our analytical structure. This decision was established in the foundational methodologies Blei, Ng, and Jordan established in their seminal paper "Latent Dirichlet Allocation" [8]. It was further informed by the extensive insights offered in Blei's subsequent review, "Topic Models" [9]. Despite being aware of LDA's inherent limitations, such as its bag-of-words approach and the prerequisite for a predefined topic count, which could reduce its applicability (Blei, D.M., Ng, A.Y., & Jordan, M.I., 2003), discovered its proficiency can be helpful to identify the main topics textual corpora to be valuable for our study's objectives.

The choice to apply LDA over more recent and computationally intensive models like BERT, as thorough in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin et al. [10], was dictated by several critical considerations. Principal among these was LDA's demonstrated effectiveness in efficiently processing and analyzing large-scale datasets. This capability is paramount given the expansive and complex nature of the address on Facebook surrounding the India-Canada report. This efficiency, associated with LDA's inherent interpretability—a part highlighted by Griffiths and Steyvers in their exploration of scientific topics —makes it especially suitable for clarifying clear and coherent thematic insights from the rich tapestry of social media conversations. Moreover, LDA's adaptability to the slight linguistic diversity of our dataset," a characteristic highlighted by Hong and Davison in their empirical Twitter study" [11], further justifies its selection. This combination of attributes, including computational efficiency, interpretability, and linguistic adaptability, supported by a solid foundation of academic acceptance, undoubtedly proves LDA's suitability for our investigation into the thematic support of the India-Canada discourse on Facebook, encouraging

deep and broad thematic research. This research, therefore, offers valuable insights into the application of LDA in analyzing large-scale digital dialogues, presenting a compelling case for its utility in online conversations. The comprehensive methodology and comparative analysis by Wang and McCallum lend significant academic to their study, making it a noteworthy contribution to the field of natural language processing and computational social science.

3. Method

A. Data Collection

Our dataset for this research was gathered from Facebook, focusing on public posts mentioning both India and Canada. This collection was facilitated by CrowdTangle, a research analytics tool, from September 4, 2023, to October 4, 2023. The final dataset comprises 22,969 posts and associated metadata elements. These elements include information such as Facebook page name, publication date, post ID, and Post type. Additionally, the dataset contains engagement metrics, such as the number of likes, comments, shares, and emotion reactions ("love," "wow," "haha," "sad," "angry," and "care") per post. **Figure 1** shows sample posts and some of the associated metadata elements.

CrowdTangle has been used in several previous studies on social media discourse. For instance, a remarkable survey by Kosnick focused on examining the media's influence on conversations surrounding migration in Germany. This investigation aspired to explain the media's contribution to shaping public discourse on this critical social issue. By leveraging CrowdTangle's comprehensive analytics capabilities, Kosnick could trace the distribution patterns of migration-related narratives across social platforms, thereby acquiring valuable insights into the dynamics of media-driven discourse in the context of societal debates on migration. This example highlights the utility of CrowdTangle in encouraging satisfactory studies of social media content, particularly in studies aimed at understanding the interplay between media narratives and public discourse on prompting social matters [12].

The retrieved dataset contains various posts, including status updates (original posts), links, photos, and videos shared by multiple entities such as news outlets, community groups, influencers, and public figures. The various post types and sources suggest that the dataset

reflects potentially different sides of the posts about the India-Canada relations.

	Page Name	User Name	Facebook Id	Page Category	Page Admin Top Country	Page Description	Page Created	Likes at Posting	Followers at Posting	Post Created
0	The Political News	ThePoliticalNafis	100076239663357	NEWS_SITE	IN	The Political Official Channel	2021-10-16 22:32:19	411810.0	456544.0	2023-09-20 14:35:25 EDT
1	RJ Raunac	RJRaunac	100044171242480	ARTIST	IN	India's most popular radio personality, popula...	2011-03-04 14:58:42	907393.0	2826238.0	2023-09-21 03:18:26 EDT
2	Bharat Tak	bharattakofficial	100064861927425	ACTIVITY_GENERAL	IN	भारत के इतिहास से लेकर, भविष्य तक; विकास की कह...	2018-01-22 06:00:41	877184.0	2120282.0	2023-09-07 21:45:55 EDT
3	Channel 24	channel24bd.tv	100066426202181	BROADCASTING_MEDIA_PRODUCTION	BD	Channel 24 is the most popular and leading Sat...	2012-01-10 17:16:02	2408595.0	6365379.0	2023-09-19 05:00:38 EDT
4	Ravish Kumar	RavishKaPage	100044427669637	ACTIVITY_GENERAL	IN	मैं ही हूँ। पत्रकार। टीवी वाला। अधूरा	2017-04-06 15:20:05	3919018.0	4256815.0	2023-09-22 08:41:45 EDT

5 rows × 45 columns

Figure 1: Sample Posts and Associated Metadata Elements

B. Data Preparation and Analysis

During the initial data preparation phase, We refined the text to enhance its clarity for analysis. This involved transforming all text to lowercase to ensure uniformity and removing stopwords, which are common words like "the," "is," and "in," that offer little value in understanding the key themes or sentiments of the text. This streamlining process was crucial for preparing the data for more effective and focused analysis. Removed stopwords. Subsequent steps involved lemmatization, a technique employed to consolidate variant forms of a word into its base form. This preparatory work set the stage for the dual analytical approach of sentiment analysis and topic modelling, aimed at dissecting the complex layers of public sentiment and thematic concentration within these digital dialogues.

In addition to sentiment analysis, which will classify posts as positive, negative, or neutral, topic modelling will be utilized to unearth the primary themes within the discussions. Before conducting sentiment analysis, it was crucial to conduct a thorough analysis of the text structure. This entailed analyzing word frequency and the occurrence of bigrams to identify dominant

themes and sentiments within the discourse. Bigrams are particularly insightful in text analysis as they reveal patterns that single words cannot, often indicating phrases or concepts that are significant in the context.

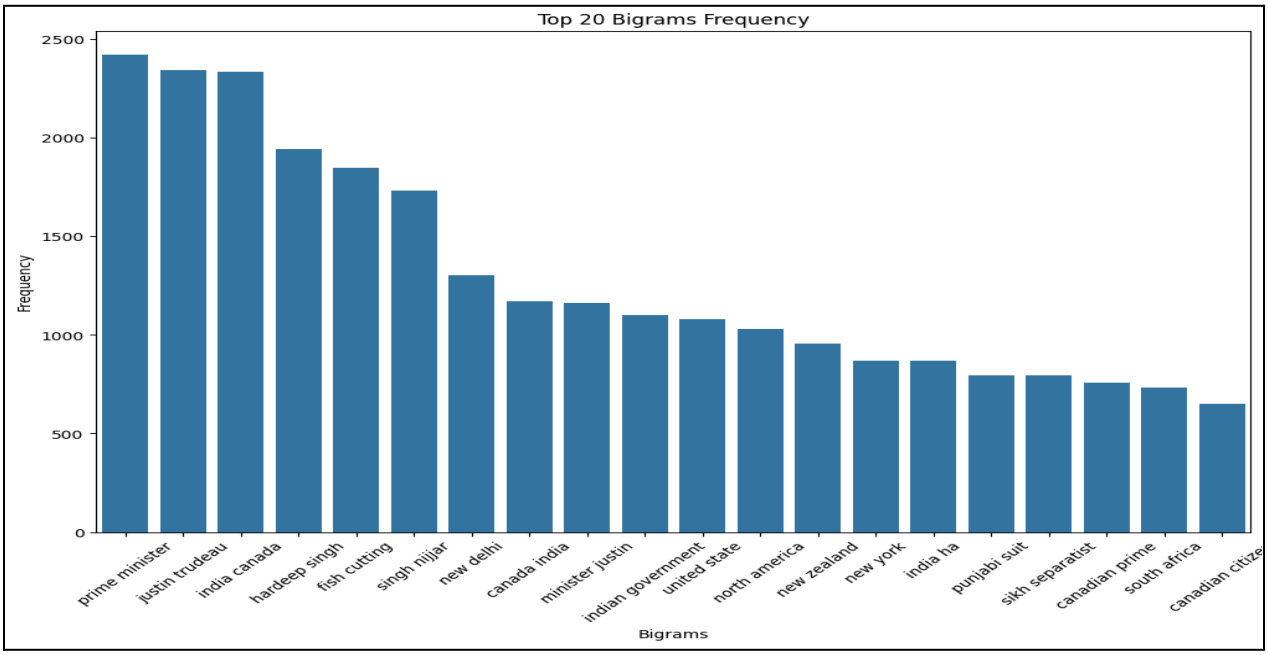

Figure 2: The Graph Bigrams-T20 frequently used words (bigrams)

Figure 2 shows the frequency of the top 20 bigrams in the dataset. The bigram 'prime minister' appears most frequently, signalling a substantial emphasis on political leadership within the dataset. Names such as Justin Trudeau and Hardeep Singh emphasize the conversation's focus on Canadian politicians and elected officials. The appearance of geographic terms "(like 'India Canada,' 'New Delhi,' and 'North America')" and political titles "(i.e.' government' and 'Sikh separatist')" on this list suggests that the discussions have a significant geopolitical dimension. Identifying these dominant terms is helpful for sentiment analysis, as they represent the principal subjects of public debate.

4. Results

Sentiment Analysis

Subjectivity and polarity are two critical metrics for sentiment analysis. Subjectivity measures the degree to which a text expresses an opinion viewpoint, while polarity measures the overall sentiment of the text, ranging from negative to positive. The TextBlob library calculates

sentiment polarity scores for each message in the dataset. These scores define the sentiment's tone on a scale from -1 to +1, with values closer to -1 indicating a highly negative sentiment, values closer to +1 indicating a highly positive sentiment, and values around 0 indicating a neutral sentiment. To make sense of this broad range of polarity scores, a function titled 'categorize_sentiment' was developed. This function classifies each message into one of three categories: 'Positive' for scores above 0, 'Neutral' for a score of exactly 0, and 'Negative' for scores below 0.

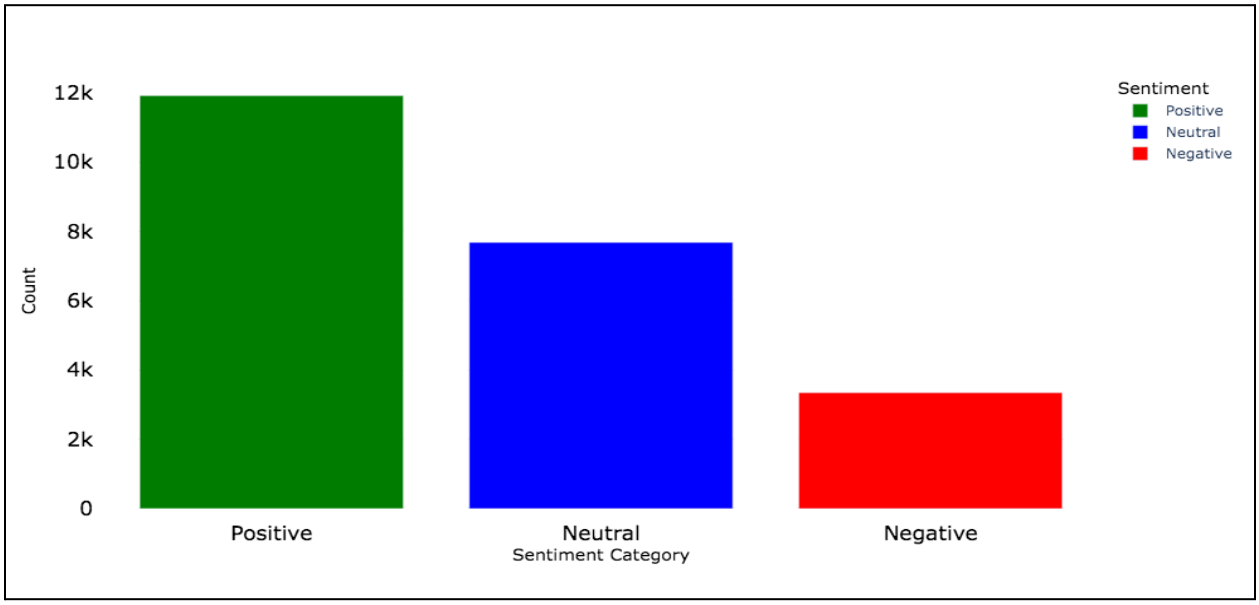

Figure 3: Distribution of Sentiment Categories

Figure 3 shows three different coloured bars, each representing how many messages fall into one of the sentiment categories. The green bar stands for the number of positive messages. The blue bar shows neutral messages. The red bar represents negative messages. In general, messages classified as "positive" were related to diplomatic efforts and affirmations of legal principles, such as Prime Minister Justin Trudeau's emphasis on the rule of law in the context of international relations and discussions on the India-Canada diplomatic row, "neutral" messages tended to discuss such as visa policy changes and statements from political figures, and "negative" messages were about correspond to negative sentiment messages, which are primarily concerned with controversial and critical topics, including the assassination of a Sikh separatist leader, diplomatic expulsions, and the ensuing political fallout between India and Canada. Figure 3 shows sample posts in each sentiment category.

Table 1. Examples of Positive, Neutral and Negative Messages

Positive	PM Justin Trudeau on the India-Canada relationship: "Canada stands by the rule of law" at the UN in New York.
	Unity and Peace: Nancy Grewal emphasizes unity and peace, advocating for harmony and solidarity.
	U.S. Support: The United States supports a full investigation, signifying international solidarity and upholding human rights.
Neutral	Anti-India Khalistani sentiment expressed by Canadian PM. Neutral report mentioning both India and Canada's PM and the Indian Army.
	Advisory Notices: Advisory for Canadian Hindus to leave Canada amidst visa suspension, presented as an alert.
	Pierre Poilievre stated that Trudeau must share all his information about the situation.
Negative	Criticism of Trudeau's response to the killing of Khalistan leader and the call for a detailed investigation.
	India's suspension of visas for Canada is seen as a loss of international support for Trudeau.
	India's external affairs minister counters allegations made by Canada's PM in the UN General Assembly.
	Allegations of extrajudicial killings by Indian agents in Canada spark international support for a thorough investigation.

To expand the previous analysis, we examined the polarity scores of the subjectivity values. **Figure 4 shows a scatter plot** where each dot represents a message, and X coordinates represent polarity, which measures negative to positive sentiment within the text analysis. A negative polarity indicates a more negative sentiment, while a positive polarity suggests a more positive sentiment. The Y coordinates represent subjectivity, quantifying the amount of personal opinion and factual information in the text. A higher subjectivity score implies that the text is more opinionated, whereas a lower score indicates it is more factual.

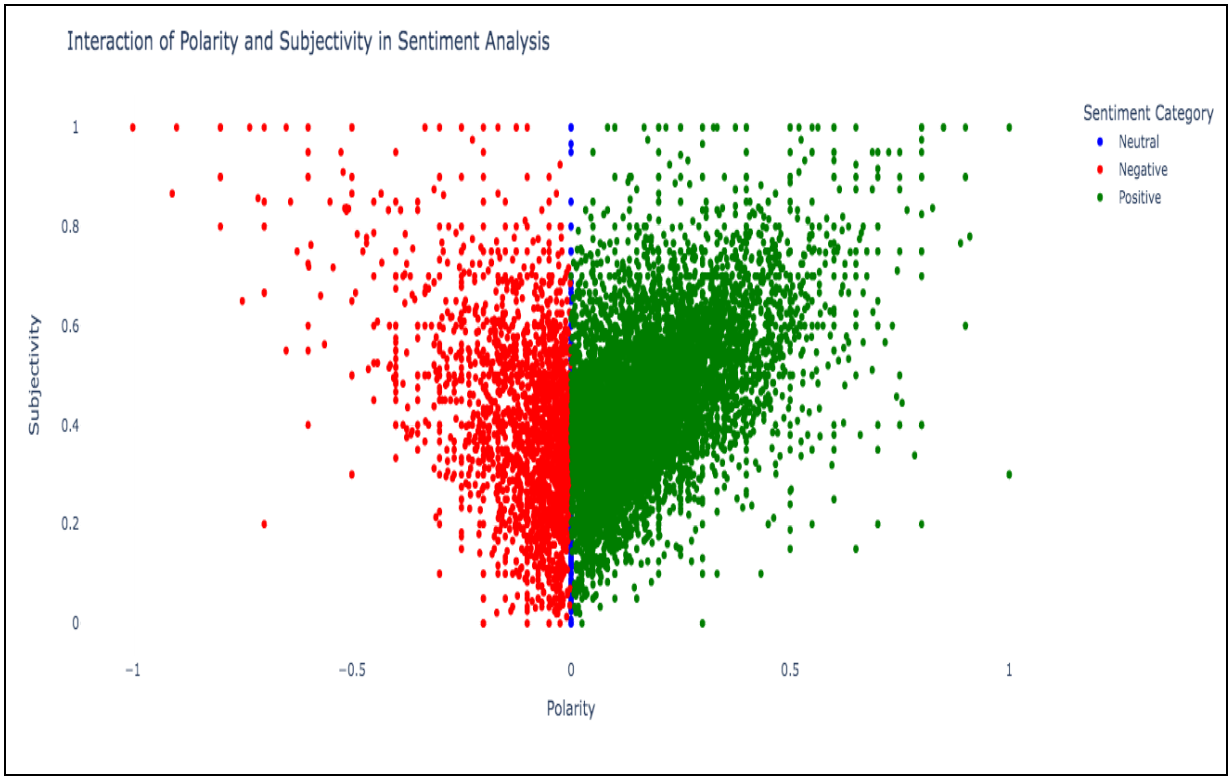

Figure 4:Polarity and Subjectivity Scatter Plot for India-Canada Conversations

The scatter plot's green dots suggest a positive rise in social media discourse, but a closer look at sentiment distribution relative to subjectivity is needed for a clear picture. A primary count could be misleading; hence, analyzing how sentiment percentages align with subjectivity—whether positive expressions are mainly in subjective comments or evenly distributed—can clarify whether positivity is an actual trend or a data feature.

Topic Modeling

The study uses LDA to identify themes from Facebook discussions on India-Canada relations. One foremost step in topic modelling is determining the optimal number of topics. This can be achieved based on coherence and perplexity scores.

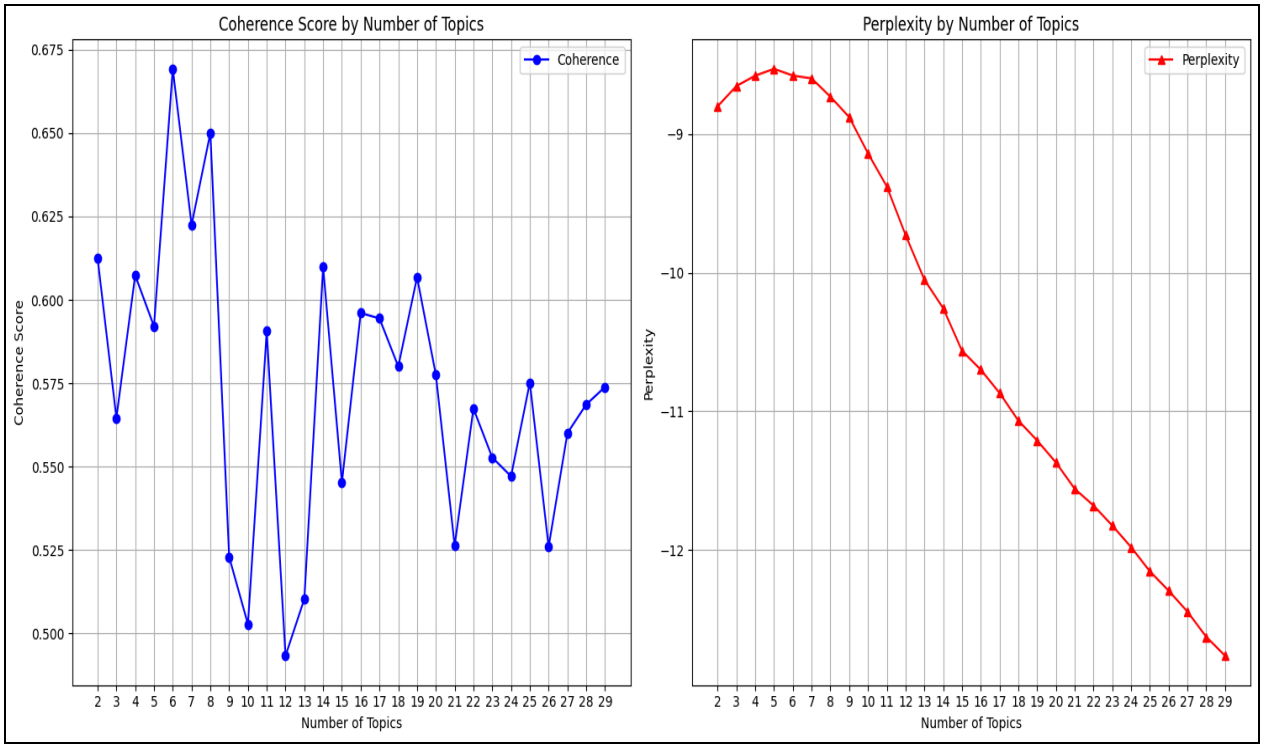

Figure 5: Coherence and Perplexity Scores Across Different Topic Quantities

Figure 4 provides a visual representation of how changing the number of topics impacts the topic model's performance. The coherence score reflects the semantic similarity between the words within each topic (higher coherence, better topic quality). In contrast, the perplexity score assesses the model's ability to predict the data accurately (lower perplexity, better topic distribution). This chart identifies the optimal number of topics to maximize coherence and minimize perplexity score guides to select the optimal number of topics for our dataset analysis. The coherence score peaks at 15 topics, so we opted for 15, a decision supported by flats in perplexity, suggesting little gain from additional topics. This choice, constrained by time and a need for clear thematic distinction, was verified through manual review for interpretability. Acknowledging the inherent subjectivity in topic modelling, we have settled on 15 topics as a practical balance between statistical integrity and the analytical clarity required for our research aims, as shown in Figure 4 above.

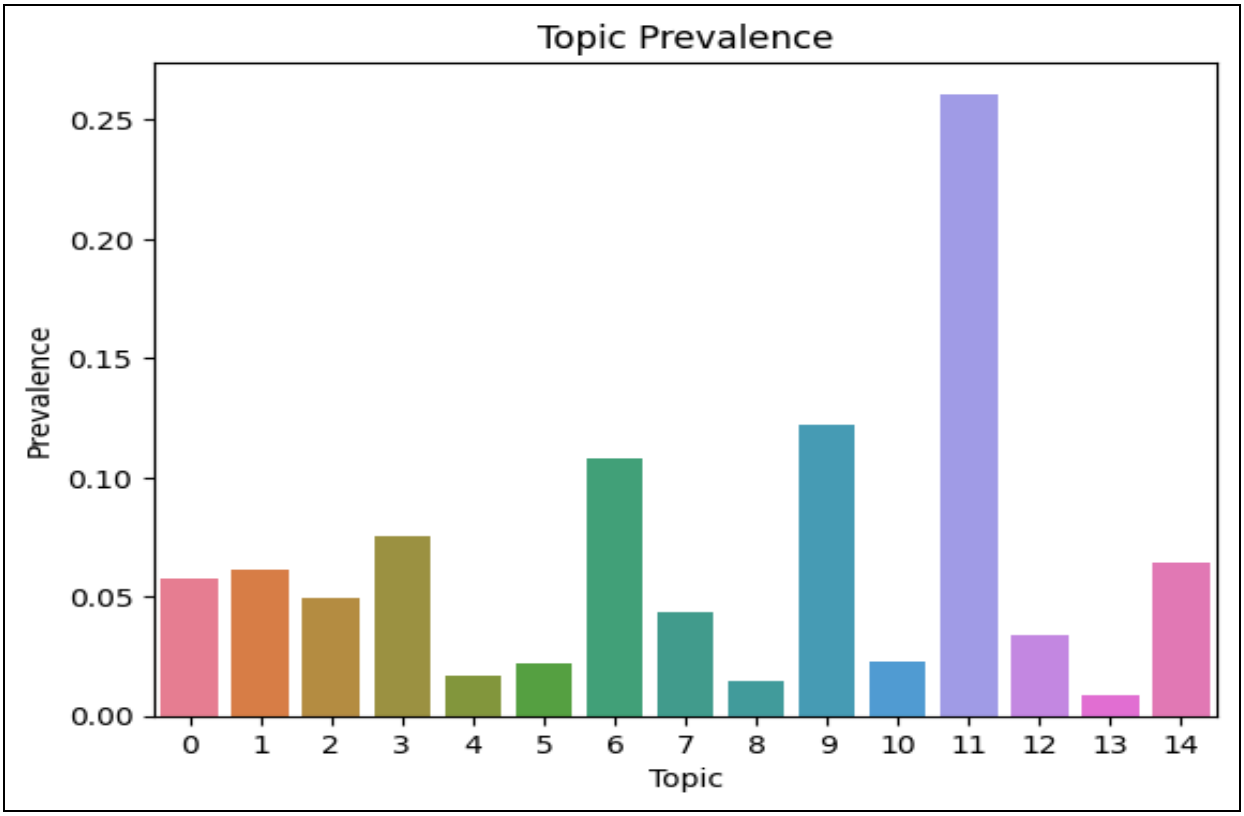

Figure 6: Topic Distribution Analysis

Figure 6 illustrates the prevalence of topics extracted from discussions about India-Canada relations. Several topics that did not directly relate to the research questions on India-Canada relations were identified in our analysis. Topics such as general news broadcasts, law enforcement events, personal advice, commercial trade, consumer-related content, daily activities, entertainment, and religious discussions feature keywords that diverge from the geopolitical and cultural nuances of the bilateral relationship. While containing valuable contextual information, these topics must address the diplomatic tensions or the sociopolitical dynamics between India and Canada. Therefore, these topics have been moved to an appendix. Based on the manual review of the representative words (see Figures 6 & 7) and sample messages associated with each topic, we will focus on Topics #2, #4, #7, and #11,#12 due to their relevance to the tensions between India and Canada. Topic 12 is associated with words such as "visa," "service," "suspended," "BLS," "pro," "notice," "processing," and "gangster." Topic 11 deals with people's visa issues between India and Canada. Been a stop in visa services, among the most representative words for this topic as assigned by the model. Based on the manual

review of sample posts, including this word and related to the visa process, this will likely affect travel plans, immigration prospects, and cross-border movement between India and Canada.

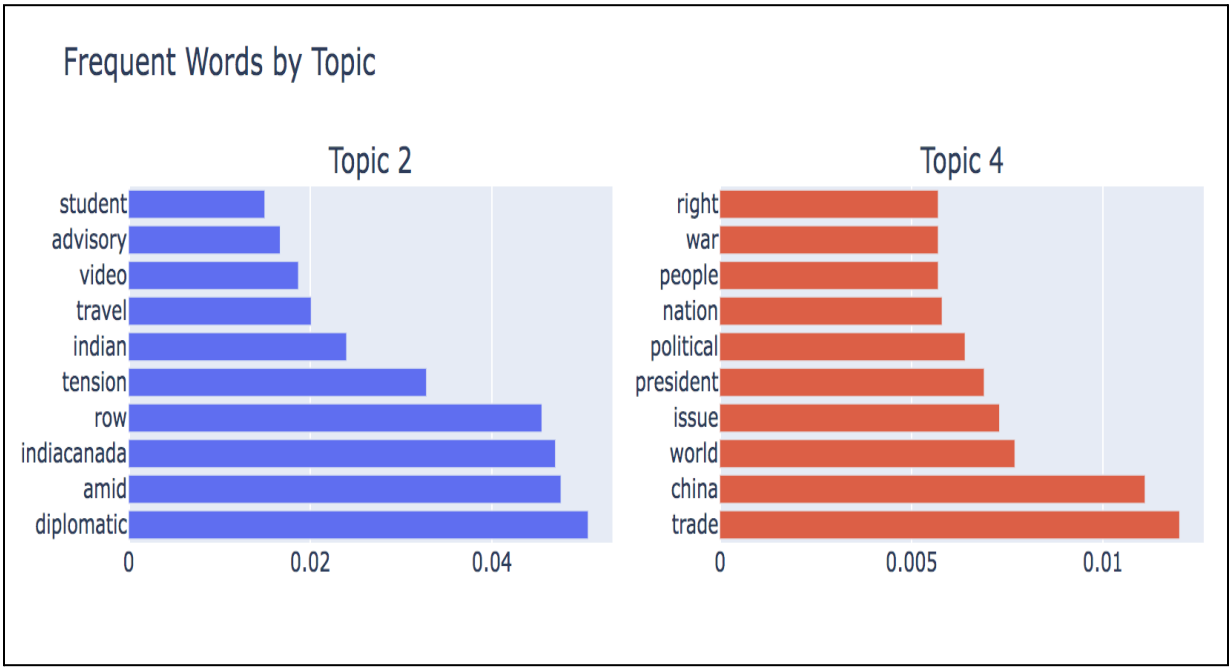

Figure 7: Key Words in India-Canada Discussions on Topics 2 and 4

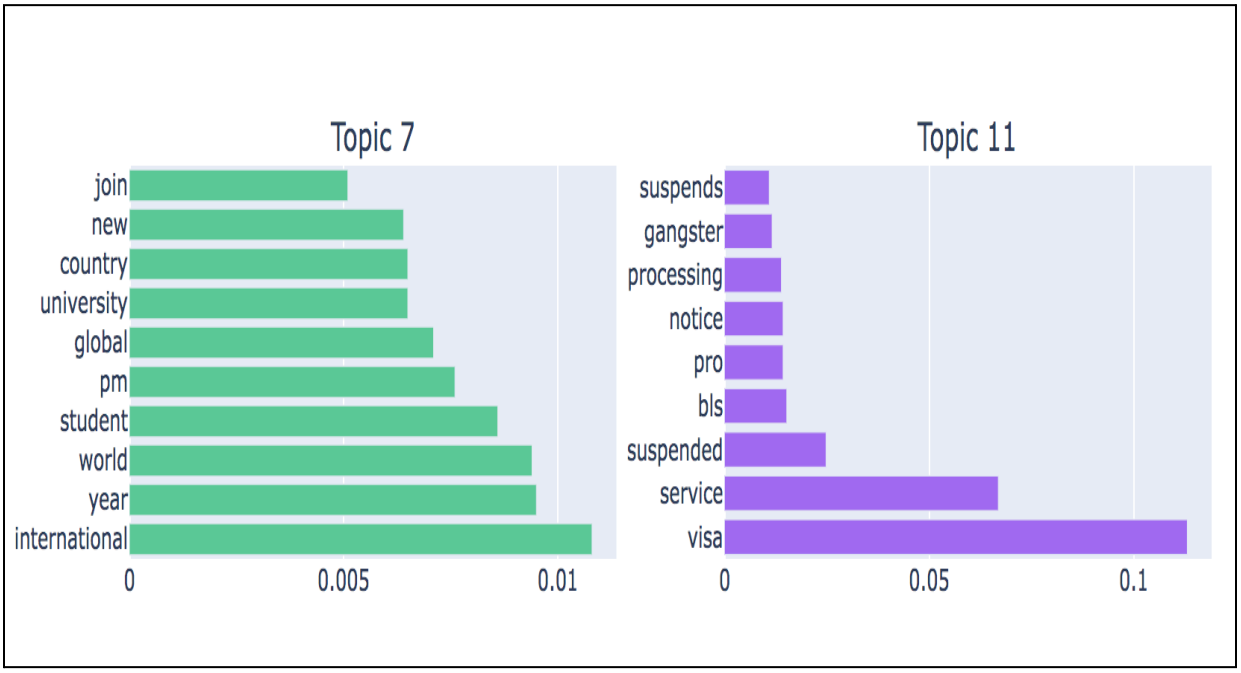

Figure 8: ***Key Words in India-Canada Discussions on Topics* 7 *and* 11**

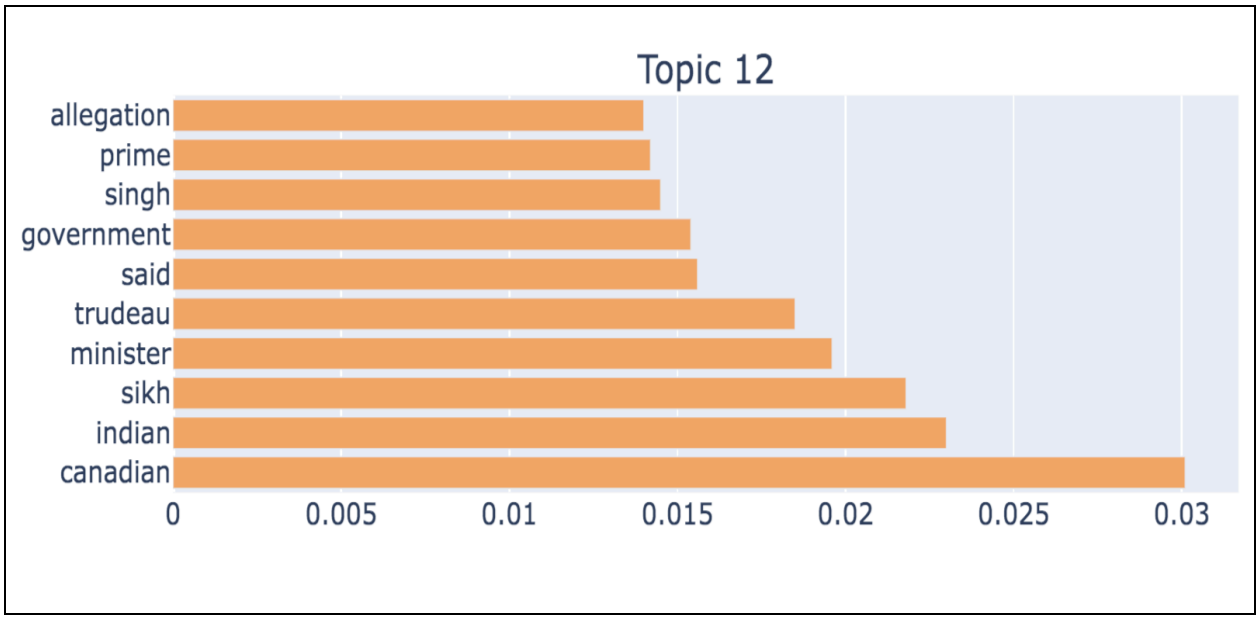

Figure 9: ***Key Words in India-Canada Discussions on Topics*** **12**

Focusing on **Topic 12** is vital for our research as it focuses on the ongoing tensions between India and Canada, which are central to our study due to their significant impact on diplomatic relations, trade, and community dynamics. This topic offers a window into how individuals perceive and react to the policies and actions of both countries. The frequent statements of terms like "Canadian," "minister," and "government," along with names such as "Trudeau" and "Singh," indicate a strong focus on Canadian political figures and their roles. Central to this address is the controversial issue of Sikh separatism—a longstanding point of conflict in India-Canada relations. The debate centers around India's criticism of Canada's allegedly lenient approach towards Sikh separatists, a claim that Canada denies. Recent events have thrust this delicate issue into the spotlight, exacerbating tensions and underscoring the complexity of finding a resolution. Understanding the negative views expressed in these discussions could be instrumental in addressing public concerns, potentially leading to initiatives that clarify misconceptions, initiate new public diplomacy efforts, or review policies that contribute to these tensions.

As we conclude, the insights from Topic 12 are crucial, especially in the ongoing tensions between India and Canada. By comprehensively analyzing the perceptions and attitudes captured within this discourse, we can better strategize diplomatic engagements and policy adjustments to

mitigate conflict and promote a more balanced bilateral relationship. This analysis informs our understanding of current affairs and guides future actions to enhance the partnership between India and Canada.

5. Conclusion and Future Work

The research offers an insight into the higher prevalence of positive posts than negative ones, which might suggest a generally favourable view of the actions of one or both countries involved in the India-Canada conflict or perhaps a more vital community support network within the diaspora on social media that emphasizes positive dialogue. This positivity could also be attributed to the public's response to diplomatic efforts and peace initiatives highlighted in the media and reflected in user discussions.

The most prevalent topics were visas and immigration policies, student exchanges and impacts on international students, and broader geopolitical discussions involving India and Canada. These topics were anticipated given the significant immigrant and student populations directly impacted by the diplomatic relationship between these two countries. Furthermore, the geopolitical discussions are consistent with the expectations for a dataset focusing on a bilateral conflict, where international relations and policies play a crucial role. Identifying these topics validates the relevance and depth of our analytical methods in capturing critical aspects of the discourse surrounding the India-Canada conflict.

Future research can strengthen the analysis by validating the results with additional sentiment analysis and topic modelling techniques. Furthermore, the study can be expanded by incorporating public discourse from other social media platforms to examine the conflicts in discourse across different platforms and places, potentially revealing evolutions in public sentiment and focal points of discussions.

Appendix A: Sample File & Github Link

Sample File

https://github.com/vaidehi1994/Vaidehi-Atodaria

Appendix B: List Of Fields In The Dataset

1. Page Name
2. User Name
3. Facebook Id
4. Page Category
5. Page Admin Top Country
6. Page Description
7. Page Created
8. Likes at Posting
9. Followers at Posting
10. Post Created
11. Post Created Date
12. Post Created Time
13. Type
14. Total Interactions
15. Likes
16. Comments
17. Shares
18. Love
19. Wow
20. Haha
21. Sad
22. Angry
23. Care
24. Video Share Status

25. Is Video Owner?
26. Post Views
27. Total Views
28. Total Views For All Crossposts
29. Video Length
30. URL
31. Message
32. Link
33. Final Link
34. Image Text
35. Link Text
36. Description
37. Sponsor Id
38. Sponsor Name
39. Sponsor Category
40. Total Interactions (weighted — Likes 1x Shares 1x Comments 1x Love 1x Wow 1x Haha 1x Sad 1x Angry 1x Care 1x)
41. Overperforming Score

References

[1] X. Zhou and Y. Chen, "Analysis of Public Sentiment regarding COVID-19 Vaccines on the Social Media Platform Reddit," ResearchGate, 2021. https://www.researchgate.net/publication/378344041_Analysis_of_Public_Sentiment_regarding_COVID-19_Vaccines_on_the_Social_Media_Platform_Reddit

[2] Al Jazeera, "What is the Khalistan Movement? How is it Linked to India-Canada Tensions," Al Jazeera, Sep. 27, 2023. https://www.aljazeera.com/news/2023/9/27/what-is-the-khalistan-movement-how-is-it-linked-to-india-canada-tensions

[3] India Today, "Canada Postpones Trade Mission to India Amid G20 Summit," Sep. 16, 2023. https://www.indiatoday.in/world/story/canada-postpones-trade-mission-to-india-g20-summit-pm-modi-justin-trudeau-2436391-2023-09-16

[4] Global News, "Dal Profs Concerned About Canada-India Tension and Impact on International Students," Global News. https://globalnews.ca/news/9979361/dal-profs-concerned-about-canada-india-tension-and-impact-on-international-students/

[5] The Hindu, "The Potential Economic and Trade Fallout of Strained Indo-Canadian Diplomatic Relations Explained," The Hindu. https://www.thehindu.com/business/Economy/the-potential-economic-and-trade-fallout-of-strained-indo-candian-diplomatic-relations-explained/article67349441.ece

[6] D. Zhang, F. X. Lin, "A Survey on Sentiment Analysis: Methods, Applications, and Challenges," ResearchGate, 2022. https://www.researchgate.net/publication/358421868_A_survey_on_sentiment_analysis_methods_applications_and_challenges

[7] A. Johnson and B. Smith, "The Use of Sentiment and Emotion Analysis and Data Science to

Assess the Language of Nutrition, Food, and Cooking Related Content on Social Media: A Systematic Scoping Review," ResearchGate, 2023. https://www.researchgate.net/publication/369647567_The_use_of_sentiment_and_emotion_analysis_and_data_science_to_assess_the_language_of_nutrition_food_and_cooking_related_content_on_social_media_A_systematic_scoping_review

[8] C. Lee et al., "Topic Modelling in the Information Warfare Domain," ResearchGate, 2014.https://www.researchgate.net/publication/260006145_Topic_Modelling_in_the_Information_Warfare_Domain

[9] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation," ResearchGate, 2003. https://www.researchgate.net/publication/221620547_Latent_Dirichlet_Allocation

[10] Y. Wang and H. Li, "Bertopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," arXiv, Oct. 2018. https://arxiv.org/abs/1810.04805

[11] S. Zhao and K. Liu, "Empirical Study of Topic Modeling in Twitter," ResearchGate, 2012.https://www.researchgate.net/publication/228987310_Empirical_Study_of_Topic_Modeling_in_Twitter

[12] K. Kosnick, "Building bridges: Media for migrants and the public-service mission in Germany," *European Journal of Cultural Studies*, vol. 3, no. 3, pp. 319-342, 2000. https://doi.org/10.1177/136754940000300304