FINAL MASTER'S THESIS

A Comparative Study of Topic Modelling Algorithms (BERTopic, IndicBERT, LSA and NMF) on an Indian Political Dataset

Author:
Shreeramyaa - 260856

Supervisor:
Prof. David Weir

September 8, 2023

Acknowledgement

I want to sincerely thank my supervisor, Prof. David Weir, for supervising this project and for sharing his knowledge with me. His valuable guidance, support, and dedicated time throughout every stage of my project have been greatly appreciated.

Finally, I want to express my appreciation to my family and friends for their continuous support and encouragement.

Abstract:

Social media sites, particularly Twitter, have lately emerged as important conversation points for Indian politics at the moment, particularly considering the recent elections. Social scientists now have new chances to analyse and evaluate these talks, including identifying community feelings and evaluating the influence of one group on another, thanks to the explosion of political speech online. It can be difficult and time-consuming to find and gather valuable information from these online forums, though. As a useful tool in this situation, Topic Modelling (TM) provides an automated method for classifying, understanding, and condensing enormous amounts of textual material. In order to do this, it finds underlying themes that are concealed inside a group of documents. Additionally, topic models deliver comprehensible representations of documents that can be used for a variety of downstream Natural Language Processing (NLP) applications. These TM methods can analyse short texts that are frequently encountered on social media platforms, handle topical correlations, and monitor the evolution of subjects over time. IndicBERT, Latent Semantic Analysis (LSA), non-negative matrix factorization (NMF), and BERTopic are the four different Topic Modelling methods that will be evaluated in this study. The strong coherence scores of BERTopic and IndicBERT, 0.71 and 0.66, respectively, show that these methods regularly produce coherent topics, according to our findings.

Table of Contents:

List of Figures:

List of Tables:

1. Introduction:

Due to the growing abundance of large datasets, there is a rising demand for TM methods to facilitate the discovery of topics [1]. TM represents a well-established computational methodology for condensing a collection of textual documents [2]. Topic models belong to a category of unsupervised machine learning methods crafted to address this objective. Their purpose lies in condensing extensive corpora of documents into precise summaries, that encapsulate the predominant themes within the corpus. These summaries manifest as topics, or clusters of interconnected words, thus earning the designation of "topic models." In precise terms, a topic model is defined as an unsupervised mathematical framework. It accepts a collection of documents denoted as D and, in turn, provides a set of topics T representing the content of D accurately and coherently.

TM holds significance as a critical undertaking across various applications. These include facet identification within sentiment analysis [3,4,5], the extraction of topics to discern user preferences in recommender systems [6], document condensation [7], the revelation of topics within chatbox systems [8], and the extraction of topics to facilitate the identification of fabricated news [9]. TM has demonstrated its efficacy across diverse domains, including but not limited to healthcare [10], online educational platforms [11], software engineering [12], and legal documentation [13], serving as a potent method for unearthing concealed thematic elements.

In this research, to comprehend the current state of politics in India, we have considered the tweets from Twitter of Indian leaders Mr. Narendra Modi (leader of the BJP), Mr. Rahul Gandhi (leader of the INC), and Mr. Arvind Kejriwal (leader of the AAP) [14]. Indian political datasets for Topic Modelling can offer in-depth insights into the political environment, public mood, and policy agendas in India. Political analysts, scholars, policymakers, and anybody else interested in comprehending the dynamics of Indian politics may find this to be helpful.

Appertaining to the insufficient knowledge of newly developed algorithms that could handle better, this study thus aims to evaluate and compare the performance of four Topic Modelling techniques, namely, LSA, NMF, BERTopic, and IndicBERT. Specifically, LSA and NMF uses a linear algebra approach for topic extraction, and BERTopic and IndicBERT use an embedding approach.

The project aims to answer the following questions:

Research Question 1:
Which of the Topic Modelling techniques, Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF), Bertopic, or IndicBERT, produces the best logically consistent topics when used on an Indian dataset?
Answer:
For all the aforementioned models, we performed Topic Modelling to determine the optimal combinations of the hyperparameters. We evaluated coherence scores using the C-V coherence technique to identify and choose the most appropriate model for the Indian dataset.

Research Question 2:
How effective are the topic labels produced by the several Topic Modelling techniques (Bertopic, IndicBERT, LSA, and NMF)?
Answer:
Following a thorough examination of the related tweets, we chose the top 10 themes and went on to manually assign labels to them. Compared the models to identify their unique advantages, disadvantages and conversations found in the dataset.

1.1 Structure of Dissertation:

An overview of the chapters of the dissertation is presented in the following

- **Chapter 1** Introduction section covers the research and outlines its objective.
- **Chapter 2** gives a comprehensive summary of previous research on the Topic Modelling.
- **Chapter 3** explains the background of the techniques utilized in this study.
- **Chapter 4** gives a detailed description of the experimental setup that has been designed to carry out the Topic Modelling, model architecture and features extracted.
- **Chapter 5** includes the findings and visualization of the results.
- **Chapter 6** provides an overview of the study, conclusions drawn from the research that has been done, and proposes future work that can be carried out for further research.

2. Literature Review:

TM algorithms have grown in popularity in twitter analysis as a useful tool for extracting insights and patterns from enormous amounts of textual material. This literature review provides a summary of significant studies and research that have investigated the application of TM algorithms in the context of twitter, such as BERTopic, IndicBERT, LSA and NMF. The purpose of this review is to examine the current state of the art and identify gaps in the literature.

2.1 BERTopic: A Transformer-Based Approach

"Analyzing Topics Discussed by Autistic Individuals on Twitter Using BERTopic [15]" examines the content and discourse patterns of tweets authored by autistic individuals. Using social media presents unique challenges and opportunities for people with autism, a neurodevelopmental condition that affects communication and social interaction. The purpose of this study is to identify and analyze the prevalent themes in tweets posted by autistic people using BERTopic, a cutting-edge Topic Modelling technique. This research aims to provide valuable insights into the interests, concerns, and communication styles of autistic Twitter users by uncovering the topics and discourse patterns within this community.

It emphasizes the importance of understanding the online discourse of autistic individuals in the discussion section of the paper. For people with autism, traditional face-to-face communication can be challenging, making social media platforms like Twitter essential

channels for expression and engagement. The discussion explores how autism manifests in online interactions and the complexities of autism.

Moreover, the discussion explores the themes and topics emerging from the analysis of tweets authored by autistic individuals. There is a high prevalence of discussions related to specific interests, advocacy for autism awareness and acceptance, and seeking social connections among online autistic communities. Moreover, the section examines the linguistic and communication patterns unique to this group, shedding light on how they use Twitter.

Research findings reveal a diverse range of topics and discourse patterns among autistic individuals on Twitter. Throughout these interviews, special interests emerge as a dominant theme, demonstrating the passionate engagement these individuals have with the subjects they are passionate about. Further, the analysis highlights autistic individuals' active participation in conversations aimed at raising autism awareness and fostering an inclusive society. Linguistic and communication patterns analysis indicate a variety of styles, including a preference for direct and literal language.

The results of this study demonstrate the power of BERTopic in identifying topics and discourse patterns within the autistic Twitter community. Results from this study not only enhance our understanding of autistic individuals' online discourse, but also hold implications for supporting their communication needs and promoting inclusivity. A deeper understanding of how autistic individuals navigate and contribute to the Twitterverse can be gained by leveraging advanced Topic Modelling techniques.

2.2 NMF: Decomposition for Topics

“An Effective Short-Text Topic Modelling with Neighborhood Assistance-Driven NMF in Twitter [16]" offers a groundbreaking solution for the intricate task of TM within the Twitter microblogging arena. The brevity and informality of tweets pose unique challenges to traditional TM approaches. A novel method based on Neighborhood Assistance-Driven NMF was developed to overcome these challenges. The primary objective of this research is to enhance the accuracy and effectiveness of TM for short texts on Twitter, resulting in improved information retrieval and a deeper understanding of content.

Twitter's increasing significance as a real-time source of information and public sentiment is discussed in the following discussion. Due to their succinct nature, tweets often lack the contextual elements needed for conventional TM algorithms to be effective. Thus, the paper critically examines existing methods and illustrates how Neighborhood Assistance-Driven NMF addresses these issues effectively. This discussion also focuses on the key components of this innovative technique, particularly the integration of neighborhood data to enhance TM precision. Additionally, practical applications such as content recommendation and trend analysis within the Twitter domain are thoroughly discussed.

The results of the research confirm the effectiveness of Neighborhood Assistance-Driven NMF for short-text TM on Twitter. The authors demonstrate through rigorous experimentation that their methodology is more accurate than traditional NMF and LDA techniques. With the inclusion of neighborhood information, the discernment of coherent and meaningful topics within concise text is significantly improved. In addition, the paper illustrates the real-world utility of this method via case studies that illustrate its prowess in identifying trending topics, summarizing discussions around specific events or hashtags, and improving Twitter content

recommendation. Overall, the results demonstrate how Neighborhood Assistance-Driven NMF can enhance our understanding and usability of short-text data on social media platforms like Twitter.

2.3 LSA: Traditional Approach

In the research paper titled "**Method of Text Summarization Using LSA and Sentence-Based Topic Modelling with BERT** [17]," an ingenious approach is presented to text summarization. Text summarization plays a crucial role in making vast textual data more accessible and usable, according to the paper. In terms of coherence and informativeness, it highlights the limitations of existing summarization techniques. By combining LSA and BERT-based sentence-level TM, the authors address these challenges. It combines LSA's ability to capture latent semantic relationships with BERT's ability to understand language to produce concise and coherent summaries. It has been demonstrated that this method produces higher-quality summaries than traditional summarization techniques using ROUGE scores and human evaluations. Research results suggest that this innovative approach could be used to improve text summarization across diverse domains, offering a powerful tool for information condensation and content extraction.

In the discussion, the paper discusses the intricacies of the proposed method. A synergy between LSA and BERT is highlighted in the step-by-step description of text summarization. The authors discuss the importance of evaluation metrics and human assessment results, which support the method's effectiveness. Additionally, the paper explores potential applications, such as document summarization, information retrieval, and content recommendation. The study highlights the transformative potential of combining LSA and BERT for text summarization, ushering in a new era of coherent, informative, and high-quality summaries.

2.4 IndicBERT: Addressing Linguistic Diversity

The paper **"IndicBERT-based Approach for Sentiment Analysis on Code-Mixed Tamil Tweets [18]"** introduces a novel approach to sentiment analysis in the context of code-mixed Tamil tweets. The practice of blending multiple languages in social media content poses unique challenges for sentiment analysis models. To address these challenges, the research uses IndicBERT, a BERT-based language model tailored to Indian languages. Using Tamil code-mixed tweets, this study examines the effectiveness of IndicBERT in improving sentiment analysis accuracy and nuance.

By highlighting the growing prevalence of code-mixing in social media content, especially in languages like Tamil, the paper initiates a comprehensive discussion. It is often difficult for traditional sentiment analysis models to accurately interpret code-mixed content because they are primarily designed for monolingual text. It is because of this limitation that advanced language models like IndicBERT are needed.

The discussion discusses IndicBERT's capabilities, emphasizing its ability to mitigate code-mixing challenges. Based on Tamil code-mixed tweets, the model performs better in terms of sentiment polarity classification. In this complex linguistic landscape, it effectively captures nuanced sentiments, outperforming traditional approaches.

An IndicBERT-based approach significantly improves sentiment analysis accuracy and precision in code-mixed Tamil tweets. Natural language processing has made substantial

advancements with the model's ability to handle code-switching and identify sentiment in multilingual contexts. This study acknowledges some limitations, including the need for larger and more diverse datasets and further exploration of fine-tuning strategies.

Overall, the results indicate that IndicBERT offers a promising solution to sentiment analysis challenges in code-mixed Tamil tweets. The study recognizes the evolving nature of code-mixing in social media and calls for continued research to further refine sentiment analysis approaches in this domain.

2.5 Gaps in existing literature:

The literature that is currently available on Topic Modelling on Twitter for political datasets shows a significant understanding and adaptation gap for these methods to multilingual and cross-cultural dimensions. Although Twitter is a global forum for political discussion, most of the research has concentrated on material in English or other frequently used languages. Given the variety of linguistic and cultural backgrounds found in political discussions on Twitter, this disparity becomes clear.

Future research should concentrate on creating culturally and linguistically appropriate subject modelling tools to close this gap. Comparative analyses within political datasets across various languages and geographies can also shed light on how subjects develop and manifest in various circumstances. By doing this, academics can acquire a more diverse and universally applicable viewpoint on Topic Modelling in Twitter political analysis.

3. Background:

3.1. Overview of Topic Modelling

In this segment, we will discuss about a complete explanation of the basic principles underlying the methods we have employed for the Topic Modelling. TM can be described as a statistical method [19],[20] that aids in uncovering hidden thematic structures present within a collection of documents, using machine learning methods [21]. In the realm of NLP, topic models are a type of generative model that provide a probabilistic structure. This structure can be understood as a method for discovering clusters of words [22],[23]. The assumption behind TM is that documents consist of a mixture of topics, each comprising a collection of words that are commonly used in the document. Through this, topics are derived by associating words that share comparable meanings and distinguishing the usage of words with many meanings Using this process, words are discovered that help define the boundaries between topics or reveal patterns within the data, leading to conclusions and decisions [19]. The purpose of this TM study is to create a topic model that uses a variety of techniques. we will apply the following techniques: BERTopic, IndicBERT, LSA and NMF [21].

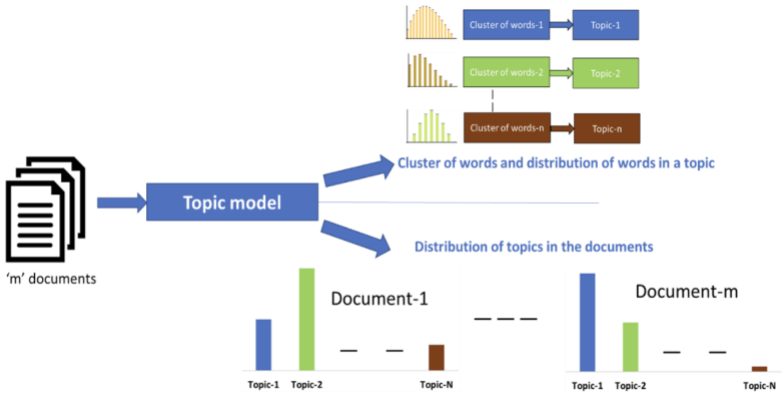


Figure 1: Topic Modelling

(Source)

3.1.1 BERTopic

BERTopic builds on Top2Vec's techniques, their algorithmic structures are comparable [24]. As an embedder, a BERT is employed, and BERTopic offers document embedding extraction in addition to a sentence-transformers model that works with more than 50 different languages. BERTopic represents a method for TM that makes use of transformers and c-TF-IDF to generate compact clusters, facilitating the creation of topics that are straightforward to comprehend. This approach also ensures that significant words are retained within the topic descriptions. The process of forming topic representations in BERTopic can be perceived as a sequence of steps. This procedure consists of five distinct stages:

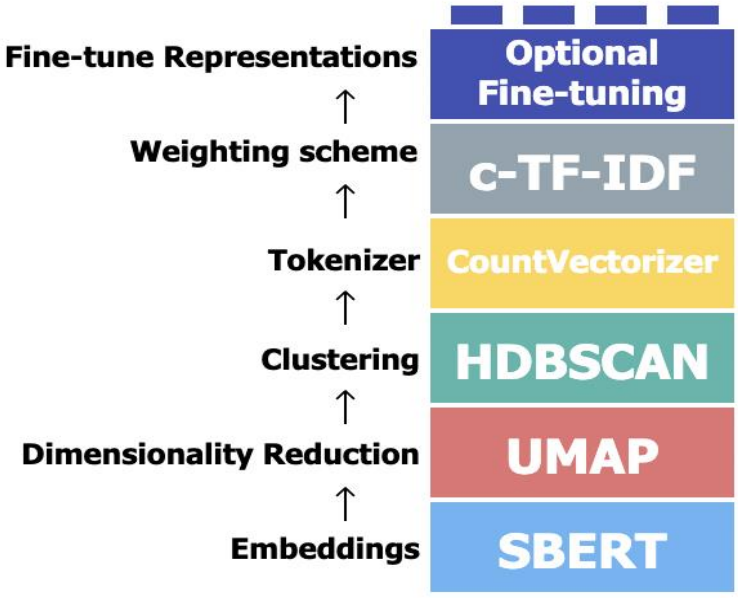


Figure 2: BERTopic Algorithm

(Source)

Although these steps are the default approach, BERTopic offers a degree of flexibility. Each step within this process was chosen with careful consideration, ensuring that they are somewhat independent from one another. For instance, the tokenization step isn't directly influenced by

the embedding model used to convert documents. This provides room for creativity in how we handle the tokenization process.

This modular effect is especially evident in the clustering step. Models like HDBSCAN assume that clusters can have different shapes and structures. Consequently, using a technique centered around centroids to model topic representations wouldn't be advantageous, as centroids might not accurately represent these diverse cluster shapes. On the other hand, a bag-of-words representation doesn't make many assumptions about cluster shapes and forms [25].

In essence, this makes BERTopic quite modular, allowing it to maintain the quality of topic generation across various sub-models. In simpler terms, BERTopic essentially empowers you to construct your own topic model according to your needs.

3.1.2 IndicBERT:

The ALBERT model (A Lite BERT for Self-Supervised Learning of Language Representations), which is a variant of BERT (Bidirectional Encoder Representations from Transformers), is where the Indic-BERT pretrained model gets its name. Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu are among the 12 Indian languages that Indic-BERT is educated in [26]. IndicBERT performs just as well as or better than other multilingual models like mBERT and XLM-R despite having less parameters. It is tested on a variety of tasks and pre-trained on a novel corpus of about 9 billion tokens. The monolingual corpus of AI4Bharat is used to pre-train the IndicBERT. The corpus's distribution of languages is as follows.

Language	as	bn	en	gu	hi	kn	
No. of Tokens	36.9M	815M	1.34B	724M	1.84B	712M	
Language	ml	mr	or	pa	ta	te	all
No. of Tokens	767M	560M	104M	814M	549M	671M	8.9B

Figure 3: IndicBERT Pre-trained Corpus

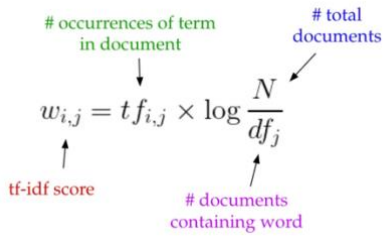
(Source)

3.1.3 LSA:

Latent Semantic Analysis, or LSA, among the fundamental methods of TM. It essentially involves taking a matrix of documents and terms and separating them into two separate matrixes: document-topics and topic-terms.

The initial step involves creating a matrix that represents the relationship between documents and words. If we have n words and m documents in our vocabulary, we can create a matrix A with m rows (one for each document) and n columns (representing words). Each matrix cell in the simplest form of LSA can simply include a count of the number of times a given word (column) appears in a given document (row). However, in real scenarios, using raw word counts doesn't work very effectively because it doesn't consider how important every word is in a document. For instance, the word "nuclear" likely provides additional meaningful information about the topic(s) of a document compared to a word like "test." As a result, LSA

models usually swap out the basic word counts in the document-term matrix (DTM) for something called Tf-idf scores. Tf-idf, which stands for term frequency-inverse document frequency, gives a weight to each word in a document like this [27]:



To break down the DTM and uncover topics, LSA uses a matrix simplification technique called Single Value Decomposition (SVD) [27]. SVD essentially splits the DTM into three separate matrices: DTM = $U \cdot \Sigma \cdot V^t$. Here, U and V are m x m and n x n in size, where m is the number of documents and n is the number of words in the corpus. Σ is m x n, with only its main diagonal holding singular values of the DTM.

In LSA, the first t (where $t \leq \min(m, n)$) largest singular values are chosen from the DTM. This means the last m - t and n - t columns of U and V are discarded. This process is called truncated SVD. The outcome is an approximation of the DTM with rank t.This approximation is optimal because it's the closest rank t matrix to the DTM concerning the L_2 norm. The remaining columns of U and V represent document-topic and word-topic relationships, with t indicating the number of topics.

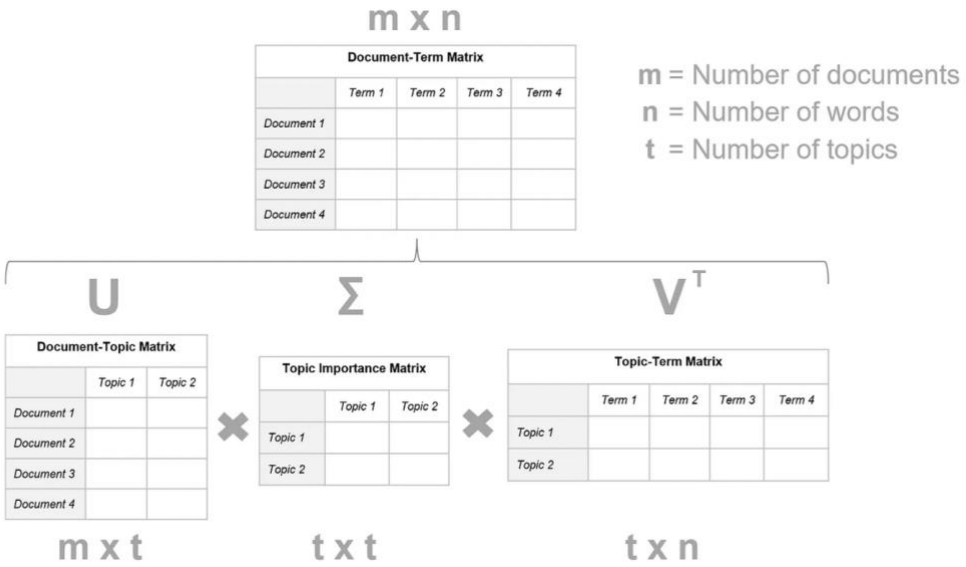


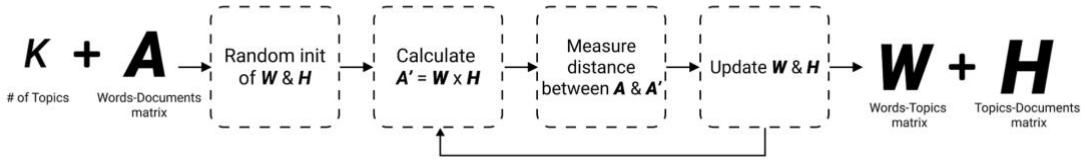
Figure 4: Truncated SVD on the Document-Term Matrix (DTM) to extract the latent variables (topics)

(Source)

3.1.4 NMF:

The NMF method is a mathematical technique that divides a matrix with non-negative values into two new matrices, where the product of the two matrices equals the original matrix. NMF is a difficult task, just like inferring generative models [28].
With regards to TM, NMF is used to factorize a large document-word matrix into two smaller matrices: the topic-word matrix and the topic-document matrix. The former represents topic

distributions over words, and topics are inferred by selecting the most significant words for each topic. The subject distribution of each document can be determined from this matrix, which represents topic distributions across documents. In this setup, the big matrix will be DxV and the smaller matrices will be kxV and Dxk, respectively (where k is the required number of subjects). This assumes that there are D documents and V words. The algorithm's is as follows:



- The input for this model is the words-documents matrix A of dimension (n\times m) and the number of topics K.

1. Randomly initialize the two matrices W of dimension (n \times K) and H of dimension (K \times m).
2. Reconstruct the matrix \hat{A} = W \times H.
3. Calculate the distance between A and \hat{A} (euclidean distance).
4. Update values in W and H based on an objective function.
5. Repeat steps 2, 3, and 4 until convergence.

- The output of the model is W and H matrices.

Figure 5: NMF Algorithm

(Source)

In summary, NMF is a way to break down a matrix into two matrices to represent topics and their relationships within a document-word context. It's used to approximate a complex problem, and its algorithm iteratively refines the matrices until convergence is achieved.

3.2 Evaluation:

A common metric for evaluating topic models is the topic coherence measure, which assesses the degree of semantic similarity between terms inside a subject. The 'c_v' coherence measure evaluates how well the words within a topic relate to each other. To calculate this coherence, each word in a topic is compared to all other topics. This comparison involves analyzing a sliding window of 110 words to determine if two words co-occur, either directly or indirectly. The process includes both direct confirmations when words appear together and indirect confirmations when words share context within the window. For each topic, the 'N' most probable words are chosen, and a 'word vector' is constructed for each word. This vector holds normalized values representing the significance of word co-occurrences. All the word vectors for a topic are then combined into a single vector that represents the overall topic. The coherence score is obtained by averaging the cosine similarities between each word in a topic and its corresponding topic vector. In essence, the 'c_v' coherence score assesses the coherence of topic words by analyzing their contextual relationships within and across topics, helping to gauge the effectiveness of the topic model. The equations 1 through 8 provided here are sourced from the article authored by Emil Rijcken [29].

We start with a corpus C with D documents and define the following quantities:

δ_d document d represented by a bag of words,
$|\delta_d|$ the number of words in document d,
$w^C_{d,i}$ corpus word at index i in document d
d document index in a corpus. $d \in \{1, 2, ..., D\}$,
i word index in document d. $i \in \{1, 2, ..., |\delta_d|\}$.

Also, we have a trained topic model with K topics and N most probable words per topic. We define the following quantities:

k topic index. $k \in \{1, 2, ..., K\}$,
n word index in a topic. $n \in \{1, 2, ..., N\}$,
W_k the set of the N most likely words in topic k,
$w^T_{n,k}$ topic word at index n in topic k,
$\overrightarrow{w}_{n,k}$ vector to represent topic word at index n in topic k.
Where $|\overrightarrow{w}_{n,k}| = N$.

(1)

A large part of the CV score is derived from the NPMI scores, an advanced method for calculating the probability of two words will appear in a corpus.

$$NPMI(w', w^*) = \frac{log\frac{P(w', w^*)+\varepsilon}{P(w')\ P(w^*)}}{-log(P(w', w^*) + \varepsilon)} \quad \textbf{(2)}$$

The formula for Normalized Pointwise Mutual Information

The epsilon serves as a minor constant that prevents the calculation of a logarithm with a value of zero. This probability computation relies on a sliding window denoted as 's' (taking a value of 110 in the case of C_v coherence). When 'j' represents the position of the sliding window within a document, the probabilities within the NPMI formula are determined in the subsequent manner:

$$P(w_n, w_m) = \frac{\sum_{d=1}^{D}\ \sum_{j=1}^{|\delta_d|-s}\ b_{d,j}(w_n, w_m)}{\sum_{d=1}^{D}\ |\delta_d| - s} \quad \textbf{(3)}$$

Where:

$$b_{d,j}(w_n, w_m) = \begin{cases} 1, & w_n, w_m \in \{w^C_{d,j}, w^C_{d,j+1}, ..., w^C_{d,j+s}\}, \\ 0, & else. \end{cases}, \ w_n, w_m \in W_k$$

The calculation of probabilities between two words, based on a sliding window s

Based on the NPMI, we generate a word vector for each topic word. This word vector has a length of 'N', and it's constructed to represent the connection of a topic word to others within the same topic (direct confirmation).

$$\overrightarrow{w}_{n,k} = NPMI(w^T_{m,k}, w^T_{n,k}), \forall m \in \{1, 2, ..., N\} \quad \textbf{(4)}$$

The creation of a word vector (direct confirmation measure)

For the segmentation, we encounter the following:

$$\{(w_n, W^*)\ | w_n \in W_k; W^* = W_k\} \quad \textbf{(5)}$$

The segmentation of word subsets

To make a comparison between each word and the topic vector, we generate K topic vectors. These vectors are formed by adding up all the N words within each topic (note that the notation W* here corresponds to the same W* mentioned below):

$$\vec{w}_k^* = \sum_{n=1}^{N} \vec{w}_{n.k}$$

The topic vector

(6)

A cosine similarity is calculated for each topic word vector based on the topic vector and segmentation. The cosine similarity is calculated as follows:

$$s_{cos}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{w}| \times |\vec{w}|}$$

The cosine similarity

(7)

The average of all NxK cosine similarities is taken to calculate the C_v score:

$$c_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{cos}(\vec{w}_{n,k}, \vec{w}_k^*)}{N \times K}$$

The C_v score is the average of all cosine similarities.

(8)

4. Methodology

4.1 Study Overview

The purpose of this study is to analyse various Topic Modelling algorithms in order to provide fresh perspectives and answers to social scientists interested in analysing political interactions. This research was motivated by the complexity of short-text social media data. In contrast to other platforms, Twitter offers succinct tweets that may be found by using particular hashtags and have a maximum of 280 characters (Queiroz, 2018) [30]. Considering user interests, the use of hashtags streamlines the information search process. Given that social media has the potential to improve crisis communication (Femenia-Serra et al., 2022) [31], this study uses Twitter posts from different leaders related to the Indian political dataset as a benchmark for assessing the four topic models mentioned above (i.e., LSA, NMF, BERTopic, and IndicBERT). Following are the specific steps that were taken to implement this study.

4.2 Experimental Environment:

The experiments were done on Visual Studio code 1.79.2, using Python 3.10.9, which is a widely used programming language due to its usability with the utilization of various integrated libraries. Several python libraries, including Matplotlib, Pandas, NumPy, seaborn and TensorFlow are used in this study to carry out the analysis and comparison of the TM. The hardware environment of the project is a Macbook Pro M1 chip operated on macOS. Achieving peak performance on macOS is made possible by the M1 Pro chip's cutting-edge 10-core CPU and 16-core GPU, which offer tremendous processing capability.

4.3 Dataset:

The dataset consists of three CSV files focusing on significant Indian political leaders: Mr. Narendra Modi (Leader of BJP), Mr. Rahul Gandhi (Leader of INC), and Mr. Arvind Kejriwal (Leader of AAP). These files each include 70,000 tweets about the relevant leader. The purpose of this dataset is to analyze the current political landscape in India, particularly with state and

parliamentary elections. Political parties may be able to better grasp their advantages and disadvantages with the aid of the dataset. Each tweet is represented in the dataset by the following columns:

- **Date**: The date when the tweet was posted.
- **User**: The username of the individual who posted the tweet.
- **Tweet**: The content of the tweet.
- **Time**: The time when the tweet was posted.

The dataset is accessible through the URL on Kaggle below

https://www.kaggle.com/datasets/soumendraprasad/201k-tweets-on-mrmodimrrahulmrkejrielecanal

4.4 Data Preprocessing:

Several text preprocessing processes are used in the data purification process using Python and several libraries, including pandas, numpy, nltk, and genism are used. Let's examine each preprocessing step individually:

- **Import Libraries:** Libraries such as pandas, numpy, matplotlib, tensorflow, seaborn, re, string, and nltk are imported.
- **Loading Data:** Load data from CSV files named "Arvind Kejriwal_data.csv," "Narendra Modi_data.csv," and "Rahul Gandhi_data.csv" into separate DataFrames named Arvind_Kejriwal, Narendra_Modi, and Rahul_Gandhi.
- **Creating Target Labels:** Add a 'target' column to each DataFrame (Arvind_Kejriwal, Narendra_Modi, Rahul_Gandhi) and assign values 0, 1, and 2 to represent each politician respectively and concatenate into a single Dataframe named df
- **Convert Date Column:** Convert the 'Date' column in the DataFrame df to the datetime type using pd.to_datetime
- **Text Cleaning:** The raw textual data was put through a thorough series of cleaning procedures. The data was preprocessed using a variety of techniques, beginning with the elimination of emojis from the text. Line breaks, hyperlinks, mentions, and other non-ASCII characters were then removed once the text was converted to lowercase. The hashtags and underscores are then removed, keeping only the words. Additionally, the deletion of "$" or "&" letters and excessive spacing
- **Duplicate removal:** Through the duplicate data removal process, any potential bias introduced by duplicate or redundant text entries is eliminated, ensuring a clean set of data for subsequent analyses. We have removed the duplicate text entries from the text_clean column and the final datasets consisted of 141,156 unique tweets.

4.5. Implementation of Topic Models:

In this section, we will evaluate the Topic Modelling of all four models, including the hyperparameter configurations we used. We will then discuss the assessment metrics for coherence that we used and outline the manual labeling process that we followed in our experimental framework.

4.5.1 Execution of BERTopic:

In this research, we utilized the BERTopic library to perform Topic Modelling on a pre-processed dataset. By using state-of-the-art techniques, we sought to identify meaningful thematic clusters within the text data. The following steps outline the experimental process:

a) Data Preparation:
The pre-processed text content has been extracted from the 'text_clean' column of the dataset df_cleaned. Lists were compiled from the resulting text data.

b) Model Initialization with Hyperparameter settings:
Using the BERTopic library, we initialized a TM model. In order to find the best hyperparameters for a Topic Modelling task using BERTopic, the algorithm performs a thorough grid search. It initialises a BERTopic model with a variety of combinations of hyperparameters, assesses the model's performance, and records the hyperparameters with the best results. The objective is to find the hyperparameter configuration that produces the most logical and insightful themes from the given text data. Using the following settings, we initialised a Topic Modelling model from the library:

- **nr_topics:** This parameter can be challenging. The number of topics that will be trimmed down after topic model training is specified. The topic model will attempt to minimise the number of topics from 100 to 20, for instance, if your subject model generates 100 topics but you have set nr_topics to 20. The cap for this study has been set at 1000.
- **min_topic_size:** The parameter min_topic_size is significant! It is used to define the smallest possible topic size. More topics are formed when this value decreases. It's possible that no topics will be formed if you set this value too high. If you lower this value, you will obtain several microclusters. In this case, the value is set to 3.
- **calculate_probabilities:** We can compute the probabilities of each topic in each document using the function calculate_probabilities. This is disabled because it is computationally expensive.

c) Model Fitting and Transformation:
The BERTopic model is trained to find theme clusters within the text data as part of the model fitting and transformation process. Each document is given topic labels that correspond to the primary themes included within the text. Topic probabilities are not computed in this instance; instead, the allocated topics are recorded in a list. The generated topic labels allow the categorization of documents according to their content and shed light on the text corpus's underlying structure.

d) Topic Visualization:
The next step is to visualise the topics that was generated using BERTopic visualisation functionality. This process resulted in a graphic representation of the themes that were uncovered. The goal was to give a clear overview of the thematic clusters within the dataset, regardless of the unique nature of the visualisation. With the help of BERTopic, we were able to use sophisticated language embeddings, particularly those produced by the BERT language model, to find latent topics in our dataset. This experimental design provided the framework for further analyses and insights that aided in achieving the manual labelling for our study.

4.5.2 Execution of IndicBERT:

The dataset, which includes English as well as a wide variety of Indian languages, relates to Indian politics. We have decided to use IndicBERT as a TM strategy in light of this language diversity. We anticipate obtaining more accurate and contextually significant clusters during the TM process by using IndicBert, which is designed to comprehend and represent a wide variety of Indian languages. Below are the experimental steps involved in the IndicBERT TM:

a) Model Initialization with Hyperparameter settings:
Using the Indic-BERT embedding model, the first stage entails initialising a BERTopic model designed specifically for multilingual text analysis. By using the same method as BERTopic, the optimum hyperparameter is found. The list of hyperparameters used by IndicBERT TM is provided below.

- language="multilingual": The multilingual model, which is obtained here, is called "paraphrase-multilingual-MiniLM-L12-v2" and covers more than 50 different languages. The model has a slightly different architecture and is trained on numerous languages, although it is extremely comparable to the base model.
- embedding_model="ai4bharat/indic-bert": the specific embedding model to be used, in this instance Indic-BERT because a significant portion of the dataset's tweets are in Indian languages.
- nr_topics=1000: The desired number of topics to be identified, set to 1000 in this experiment.
- calculate_probabilities=False: This parameter indicates that topic probabilities will not be calculated for documents.
- min_topic_size=3: The minimum number of documents a topic must contain to be considered, set to 3 in this case.

b) Fitting and Transforming Data:
The fit_transform method is used to fit the initialised modelindic to the preprocessed text data in the following phase. Based on the predictions of the model, this step assigns subject labels to each document. The given topic labels are held in the resulting indictopics list, and topic probabilities, if calculated (which they are not in this case), would be held in probs.

c) Topic Visualization:
In the code, the visualize_topics method is used to visualize the discovered topics after the model is fitted and topics are assigned to the documents. In this visualization, you can see how documents are distributed across topics, giving you a better idea of what the dominant themes are within the dataset.

4.5.3 Execution of LSA:

This section outlines the procedures for using LSA to analyse the text dataset for Topic Modelling using dimensionality reduction in text data.

a) TF-IDF Vectorization:
The text data is vectorized using TF-IDF to create a numerical format appropriate for LSA. The TF-IDF values display the weighted relevance of each word in each document in relation to the corpus. To process data quickly and extract relevant information from corpora, conversion to numerical representation is crucial.During initialization, we can regulate the transformation's behaviour by setting specific parameters. After testing the default settings and the below customised Vectorizer's settings, we decided to go with the latter as it yields a more meaningful and efficient representation of the data.

- **max_df (Maximum Document Frequency):** A high percentage of words in a document can be considered common and uninformative. Max_df determines the maximum frequency each word must appear in documents in order to be included in the vocabulary. Max_df=0.95 means that words appearing in 95% or more of documents will not be included in vocabulary.
- **min_df (Minimum Document Frequency):** Words appears in very less documents contributes purposeful to topics. The min_df parameter sets the threshold for the minimum document frequency a word must have to be included in the vocabulary. In this case, min_df=2 means words that appear in fewer than 2 documents will be exempted.
- **max_features:** By setting the max_features parameter, the vocabulary is limited to a certain number of words that are most frequently used. In this scenario max_features=1000, the vocabulary will only include the top 1000 TF-IDF words.
- **stop_words:** The stop_words parameter mentions a stop words list to be excluded from the text. 'english' is a predefined list of English stop words.

b) The corpus creation:
For LSA compatibility, the TF-IDF matrix is converted into a Gensim corpus. The TF-IDF matrix is transposed and converted to the Gensim format to build the corpus. This will help in the effectiveness of the memory usage and integration with other NLP tasks.

c) Hyperparameter settings for LSA model:
LSA model is executed using the Gensim library's LsiModel that helps to understand and group words with similar contextual meaning in a large dataset. The primary hyperparameter used here is the number of topics to extract from the data. A greater number of topics may be able to capture more subtle distinctions, while an excessive number of topics may provide overfitting or less comprehensible findings. After several tests on topics counts were conducted, the value of 1000 num_topics was decided upon since it produced superior results to the others.

d) Extracting Topics and Top Words:
The topics and top words related with each subject are extracted using the show_topics technique of the LSA model, with a limit of 10 for optimal viewing.

e) Visualization with Word Clouds:
To represent the most important words within a topic visually, word clouds are created for each topic. As the best and most popular method for visualising the LSA model, the Word Cloud library is used in this visualisation.

4.5.4 Execution of NMF:

In linear algebra, NMF is a decompositional, non-probabilistic algorithm that uses matrix factorization (Egger, 2022b) [32]. NMF works on TF-IDF transformed data by splitting a matrix into two lower-ranking matrices (Obadimu et al., 2019) [33]. The TF-IDF is a measure that assesses the importance of words within a document collection. In this section, we describe the experimental setup used to accomplish NMF for TM.

a) TF-IDF Vectorization:
The text data is vectorized using TF-IDF to create a numerical format appropriate for NMF similar to LSA model. The TF-IDF vectorizer (tfidf_vectorizer) is created with the below parameters

- **max_df:** Specifies the maximum document frequency for a word to be included in the TF-IDF calculation. Words appearing in more than 85% of the documents are ignored.
- **min_df:** Specifies the minimum document frequency for a word to be included in the TF-IDF calculation. Words appearing in fewer than 5 documents are ignored.
- **stop_words:** English stopwords are removed during the TF-IDF calculation.
  The tfidf_matrix is created by fitting the TF-IDF vectorizer to the text_data. This matrix represents the numerical representation of the textual data based on TF-IDF values.

b) Hyperparameter setting for NMF:
The number of distinct themes or topics that the Topic Modelling algorithm seeks to find within the text corpus is crucially influenced by the num_topics hyperparameter. This option essentially specifies the level of classification for the textual material that the algorithm performs. The right num_topics value must be chosen carefully, taking into account factors like data qualities, study goals, and human interpretability. Setting it too low could result in issues that are oversimplified and miss small differences in content. In this case, the value is set to 1000 after conducting a multiple assessment.

c) NMF model (nmf_model):
It builds a reproducible instance of the NMF model using a fixed random seed (random_state) and a predetermined number of topics (num_topics). To identify the required number of themes in the text input, this model is trained on the TF-IDF matrix.

d) Keywords for Each Topic and Visualization:
The code extracts the indices of the top terms depending on their contribution to the topic after training the model. Using argsort(), the top words are determined, and the most important words are then retrieved. Each topic is visualised using a word cloud to depict the most significant words inside it.

4.6. Coherence

To evaluate the performance of the topic models, there is no agreed standard metric; however, coherence is used to justify its performance. In TM techniques, topic coherence is a common metric for evaluation, and the higher the score, the better [34]. Among human judgements of topics, Röder et al. [35] found the coherence score to have the highest correlation. The below explains the methodology used.

4.6.1 C_V Coherence for BERTopic and IndicBERT:

As part of pre-processing, the code organizes the text documents according to their assigned topics first. After that, it extracts the vectorizer and analyzer from a BERTopic model, enabling the extraction of feature names and terms associated with each topic and ensuring meaningful evaluation by filtering out empty topics.

The coherence score of the subjects is determined by the Coherence Model class from the Gensim library. It aids in assessing the usefulness and quality of the topics produced by a topic model. The following parameters are used to initialise the coherence model:

a) Topics: A list of each topic's key words. Each item on the list is a list of words that describe a particular subject.
b) Texts: The Bag of words representation of the pre-processed text documents for each topic.
c) Corpus: The text documents' BoW representation. It is a list where each element is a list of (word_id, frequency) tuples that each represent a document.
d) Dictionary: The dictionary object that maps words to IDs.
e) Coherence: coherence scores are derived using a specified measure ('c_v' in this case).

According to the coherence model, coherence scores are derived using a specified measure ('c_v' in this case). In 'c_v' coherence, the coherence between all pairs of words in a topic is calculated. Using the coherence measure, we can determine how often these word pairs appear together from the entire corpus versus their individual frequencies. We calculate an overall coherence score by averaging the coherence scores calculated across all topics. To obtain an overall coherence score, individual coherence scores are summed and averaged.

4.6.2 C_V Coherence for LSA and NMF:

For calculating the coherence score, a tokenized text data is used. Using the Gensim dictionary words are mapped to unique integer IDs. It is essential for further processing to have access to this dictionary. We convert the tokenized data into a Gensim corpus where each document consists of (word_id, word_frequency) pairs. Coherence is calculated by using the Gensim dictionary and ‘c_v’ coherence measure is used as inputs to the coherence model by passing the topics generated by LSA and NMF.

4.7 Topic Labelling:

We begin by gathering the labels that the models use the most frequently. These labels provide insightful cues regarding possible groupings or topics in our dataset. This first step is crucial because it establishes the framework for organising and understanding the information in our dataset. The model then extracts the subjects connected to each of these themes. These subjects are presented as lists of words or phrases, effectively capturing the essence of the information

associated with each label. We gain important knowledge about the themes that each cluster or label is focused on as a result of this extraction procedure. We organise and show the extracted subjects for each label using the top three keywords for each subject to ensure clarity.

4.8 Manual Labelling:

It is crucial to manually classify the top 10 topics produced by a model in a number of disciplines, including information retrieval, machine learning, and natural language processing. The topics produced by the models are reviewed, compared and given meaningful labels by us in this phase. The top 10 topics generated by all the models using the following procedure:

4.8.1 Manual Labelling for BERTopic and IndicBERT:

To do manual analysis on a group of tweets using the results of a BERTopic and IndicBERT model, we constructed a function called "thematic_analysis." To accommodate twitter content and its accompanying subjects, the method builds a structured pandas DataFrame, with columns labelled "Tweets" and "Topic." Then, using the TM approach, it finds the distinct themes in the dataset and goes on to map those topics to their corresponding keywords. This mapping is essential for clarifying each tweet's thematic focus. The function further enhances the DataFrame by adding a "Keywords" column that is filled with the keywords related to the subjects of each tweet. In addition, lists of words are always included in the "Keywords" column, regardless of the topic's specificity.

4.8.2 Manual Labelling for LSA and NMF:

- DataFrame is created with the following columns:"TopicNum," "Keywords," and "Sentence."
- The data is effectively clustered because Topic number is assigned based on the greatest value in each row of a matrix for Concatenating the pertinent keywords from a list that has already been defined based on the specified cluster produces the term "keywords."
- The code then determines each cluster's size and records it in a brand-new DataFrame called cluster_sizes. In order to include cluster size information in the original DataFrame, the two DataFrames are then combined based on the 'TopicNum' column. The DataFrame is finally sorted in descending order of cluster size, enabling a perceptive examination of the distribution of the data across various clusters or subjects.

Following the determination of the Top 10 topics, we thoroughly assessed the tweets connected to each of those subjects and compared them with the created keywords. Then, we carefully examined the tweets and provided descriptive labels that succinctly captured the main ideas or issues covered within each topic. This was done as part of a comprehensive manual labelling procedure.

5. Results:

A summary of the results obtained from the proposed models is presented in this section, based on various types of features described in chapter 4. Along with the topic labelling findings for each of these models, this will also contain an evaluation of their coherence score. We will also compare the labels that were manually assigned to each of the models below.

5.1. Topic Modelling:

5.1.1 Results for BERTopic:

Utilising visualisations for BERTopic and its related components is crucial to gaining a thorough knowledge of the model, how it works, and most importantly, how it can be used. The visualize_topics() method, which creates a two-dimensional representation of the topics, can be used as a practical solution. By showing how closely related the topics are to one another, this visualisation sheds light on their connections. In the visualisation, the proximity of two topics shows how related they are. There have been 980 topics generated in this specific case. The following are some observations regarding the outcomes for the below figure 6:

- The most common topic, Topic 0, is situated in the middle of the map. This shows that it is a broad subject with several connections.
- Topic 140 can be found in the top-left corner of the map and is largely used to debate the death of famed actor Sushant Singh Rajput and the pursuit of justice in that case. With few connections to other topics on the map, this placement suggests that this topic stands out as unique.
- The topics 530,17,664,8,17,981,912 and793 are all grouped together in the map's lower-right corner. The implication is that they are related to one another. The tweets give a thorough understanding of the Indian National Congress (INC) party's dynamics and leadership. They draw attention to a time of unpredictability and change inside the party, during which some party members voiced their displeasure with the Gandhi family's direction while others supported Rahul Gandhi's "Bharat Jodo Yatra." The INC is also getting ready for a forthcoming manifesto unveiling, an important occasion in the run-up to elections. A notable example of prospective changes and realignments in the political landscape is the rumour that Ghulam Nabi Azad, a former chief minister of Jammu and Kashmir, is considering starting a new political party. These tweets provide insightful information about the complex realm of Indian politics at the time.

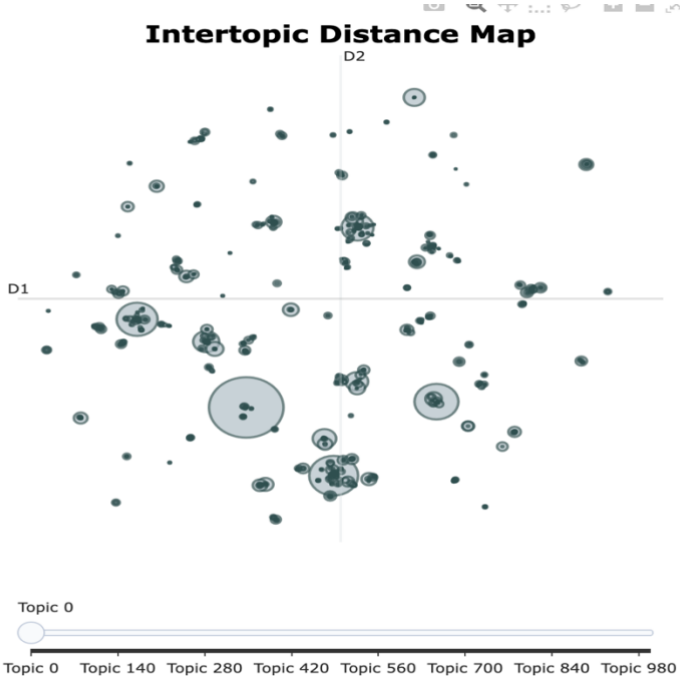


Figure 6: BERTopic's Intertopic distance map

Topic Word Scores evaluate the quality and coherence of individual subjects, which is a critical component of Topic Modelling. These ratings, which range from 0 to 1, are how well words within a subject correspond with its related theme. We used these scores in our analysis of the top 8 topics. A more unified theme is indicated by higher ratings, which reflect a greater thematic alignment. The topic word scores in this instance are comparatively high, which shows that the topics are coherent.
The tweets offered express a variety of political viewpoints. It expresses their respect and support for Rahul Gandhi, while others wish Prime Minister Narendra Modi a happy birthday. Arvind Kejriwal and various perspectives on him are discussed in tweets, while the Bharatiya Janata Party (BJP) and its leaders are also mentioned. There are also tweets attacking Arvind Kejriwal, some of which are addressing the controversy surrounding Rahul Gandhi's "Bharat Jodo Yatra." These tweets exhibit the varied and frequently divisive character of Indian political discourse and public opinion. The tweets on Rahul Gandhi's Yatra showed the highest level of coherence out of all of them in comparison to the rest.

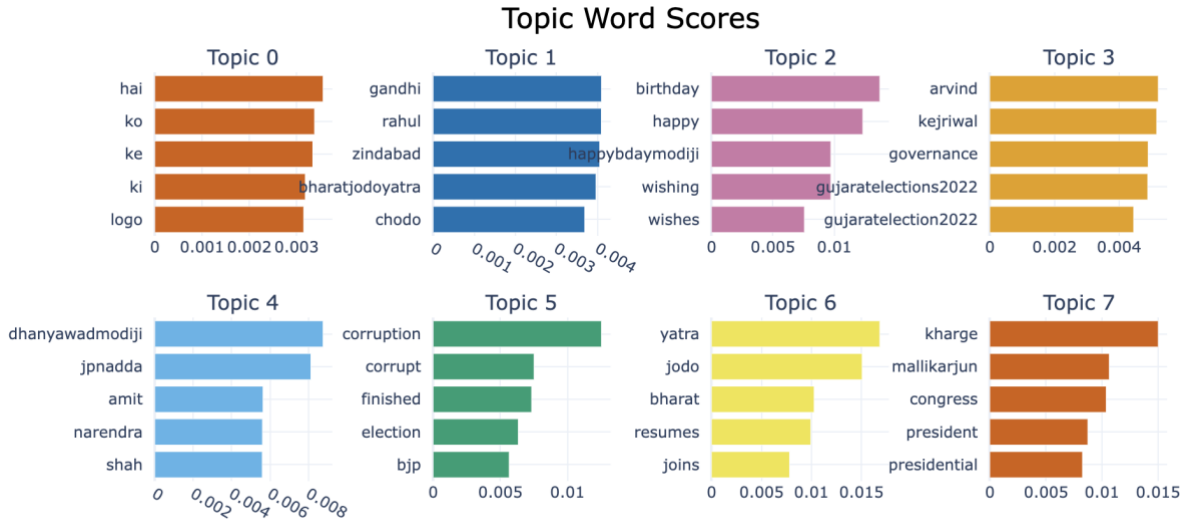


Figure 7: BERTopics's Topic word Scores

The quantity of created topics has a big impact on how well subjects are represented in Topic Modelling. It's critical to identify topics that can be merged because the separation of other topics depends on this choice. In order to shed light on potential hierarchical links and gain understanding of potential subtopics within the dataset, hierarchical Topic Modelling is used.
In this, we are focusing on the top 50 topics and the detailed explanation of the clusters listed below

- **Cluster (Red):** The tweets, which include birthday greetings for Prime Minister Narendra Modi and support messages for politicians like Rahul Gandhi, demonstrate the varied and lively dialogue around Indian politics. They offer a view of how the Indian people interacts with politicians and the government.
- **Cluster (Yellow):** These clusters primarily talk about the contacts between Prime Minister Narendra Modi and the war in Ukraine and his attendance at foreign summits. They also emphasise birthday greetings for the Prime Minister, underlining the importance of his position in Indian politics.
- **Cluster (Turquoise Blue):** The criticism of Rahul Gandhi, his public persona, and his actions, as well as political disputes involving individuals like Anna Hazare and Arvind

Kejriwal, purported scams, and the inaugurations of Prime Minister Narendra Modi's projects in Ujjain and Surat are just a few of the themes covered by these clusters.

- **Cluster (Green):** These clusters of tweets cover a range of subjects, including claims of Rahul Gandhi's fakeness, controversies surrounding his image, and his political actions, such as his participation in the Bharat Jodo Yatra. There are also tweets on Arvind Kejriwal, his alleged lies, and the controversy surrounding his Aam Aadmi Party, which includes claims of attacking Hindu deities and involvement in the Khalistan issue.
- **Cluster 4 (Purple):** The tweet collections include a variety of subjects, such as women's support for Rahul Gandhi's initiatives, Arvind Kejriwal's campaign promises in Gujarat, and young support for Rahul Gandhi's Bharat Jodo Yatra. There are also tweets honouring Narendra Modi's stint as prime minister and noting his role in the rollout of 5G services and the release of cheetahs into Kuno National Park.

The Hierarchical clustering indicates clear thematic relationships between the groups. A thematic correlation talks about Prime Minister Narendra Modi's leadership and the larger political landscape is first suggested by Clusters (Red) and (Yellow), which focus on common themes relating to Indian politics and the crucial role played by him. Similar to Cluster (Turquoise Blue) and Cluster (Green), which explore the complexities of political disputes and acts and prominently include individuals like Rahul Gandhi and Arvind Kejriwal, respectively, Cluster (Turquoise Blue) and Cluster (Green) also highlight a thematic linkage within the field of political issues. Additionally, Cluster (Yellow) and Cluster 4 (Purple) both illuminate Narendra Modi's leadership and his numerous projects, revealing a thematic thread that emphasises on his position and successes. The final two clusters, Cluster 4 (Purple) and Cluster (Turquoise Blue), investigate aspects of Narendra Modi's administration and efforts, implying a thematic relationship including his administration's acts and project implementations.

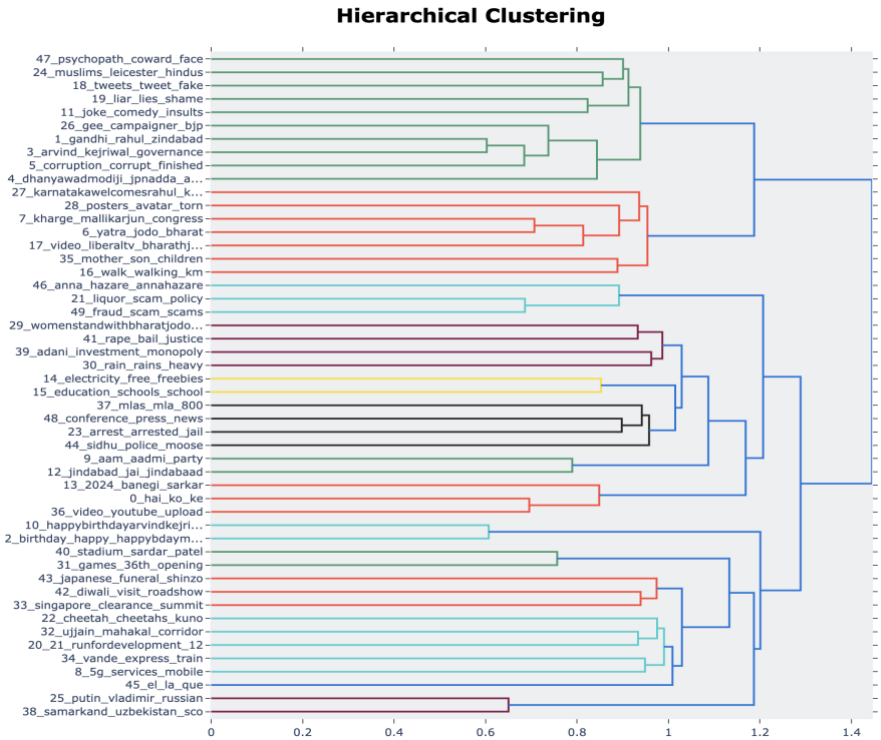


Figure 8: BERTopics's Hierarchical Clustering

5.1.2 Results for IndicBERT:

We selected IndicBERT because it has expertise working with Indian languages and is adept at comprehending their distinctive characteristics. Due to the fact that our dataset relates to Indian politics, this choice was made. Considering how well the Indian language functions, IndicBERT created larger, more significant clusters when compared to BERTopic. The largest cluster, which is in the upper left corner of the map, is made up of the most common topics (for instance, Topics 0, 972, and 50). This signifies that the subject is broad and connected to a variety of different subjects.

Several clusters have formed in the lower left corner, with Topics 21, 15, 57, 138, 81, 42, 2, 891, 265 and many more being the most prominent one. The tweets cover a wide range of subjects pertaining to Indian politics and government efforts. As Chief Minister Arvind Kejriwal prepares to open many schools, the first batch of tweets focuses on Tamil Nadu's adoption of the Delhi education model. The 36th National Games' opening ceremony, which took place at Ahmedabad's Narendra Modi Stadium, is highlighted in the second set. Prime Minister Narendra Modi was there. The final set talks about slanderous comments made about Rahul Gandhi on social media and mentions an apology and a legal notification. The expenditure of money by Arvind Kejriwal on advertisements is questioned in the fourth series. The last set highlights a disagreement between Arvind Kejriwal and Anna Hazare about Delhi's liquor laws and the placement of flags in the city under Kejriwal's tenure as chief minister. These tweets are a reflection of a range of political debates, governmental acts, and common beliefs.

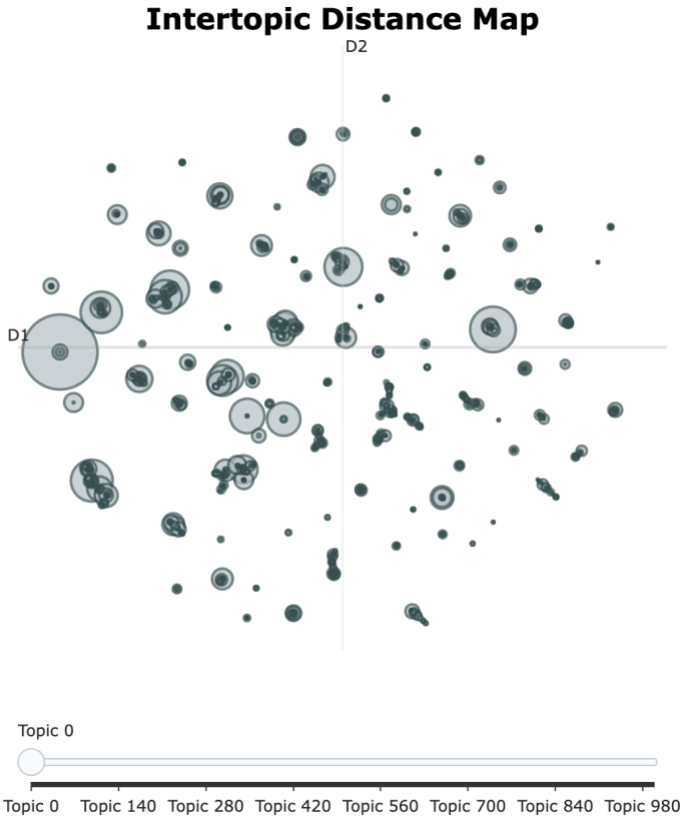


Figure 9: IndicBERT's Intertopic Distance Map

In Figure 10, Key theme categories were identified by looking at the top 5 terms connected to each of these topics. Notably, a "liquor cartel" in Punjab is one of the main issues raised in the

dataset, raising the possibility that there are flaws with the way the liquor sector operates. The study also turned up comments about "Rahul Gandhi's yatra," indicating a focus on this well-known figure's political campaigning efforts. Additionally, themes relating to the Muslim and Hindu populations appeared, indicating debates covering a broad range of social, political, or cultural issues pertaining to these religious communities. These results throw light on the most important issues of discussion by giving a succinct but insightful review of the major themes in the dataset.

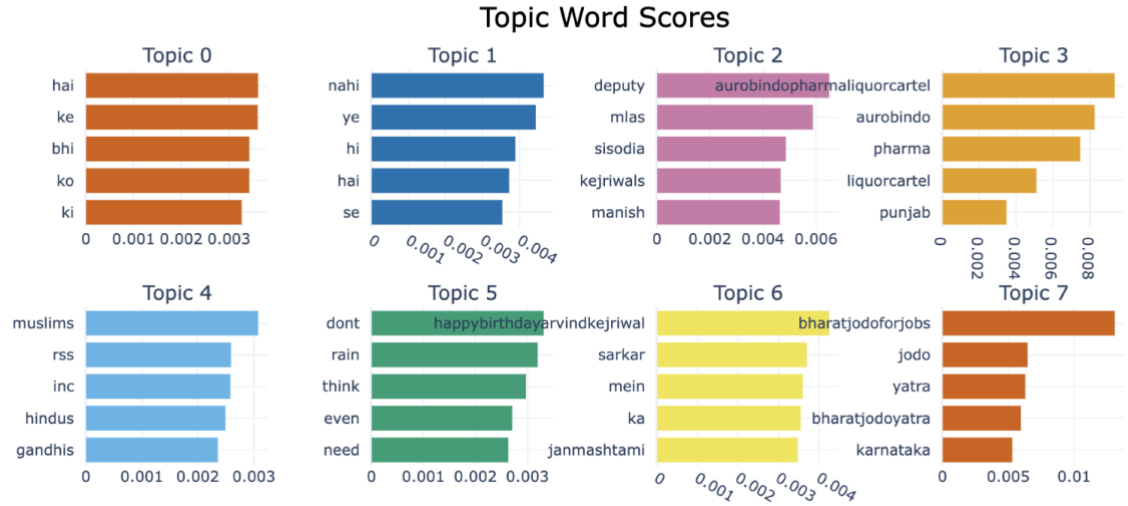


Figure 10: IndicBERT's Topic Word Scores

Comparative study of BERTopic and IndicBERT Topic word scores:
Two distinct Topic Modelling algorithms generated the subjects in the sample text. Both methods seem to produce quite similar topic results, despite some changes. Each algorithm extracts information on political figures and events in India. For instance, themes like "Rahul Gandhi," "Narendra Modi's birthday," "Arvind Kejriwal," and "Hindu-related controversies" can be found in both groups of topics. This consistency in topic extraction implies that the primary themes in the text data provided, which are centred on Indian politics and political leaders, are successfully captured by both IndicBERTopic and BERTopic. The algorithm's suitability for a certain use case may require additional analysis, such as topic coherence or relevance to particular goals.

5.1.3 Results for LSA:

We performed TM technique using LSA model for 1000 topics. We then processed the data to create a graphic of a word cloud that emphasises the top 9 topics. The prevalence and importance of numerous political events that have taken place in the world of Indian politics are demonstrated through this visual representation.

The top 10 terms connected with each topic in this word cloud visualisation clearly indicate the prevalence of particular themes. It becomes clear that the Gujarat elections and the influence of the Aam Aadmi Party on these elections are given a lot of attention. Birthday celebrations for Indian leaders Modi and Arvind Kejriwal are also apparent. Additionally, we explore Topic 6, which focuses on Manish Sisodia, an Indian politician and former social activist, and provides insights into this particular facet of Indian politics.

Figure 11: LSA Word Cloud

5.1.4 Results for NMF:

In this analysis, we display word cloud visualisations for the top 5 subjects, highlighting the 10 most frequent words within each topic. In contrast to other TM techniques, it's important to note that the subjects produced by this specific model are somewhat ambiguous. Two main themes are the main areas of discussion in these issues: the first is connected to the Prime Minister and Arvind's birthday celebration and the controversies surrounding these events. The second issue includes comments about the interactions between Congress party officials and their electoral campaigning in Tamil Nadu.

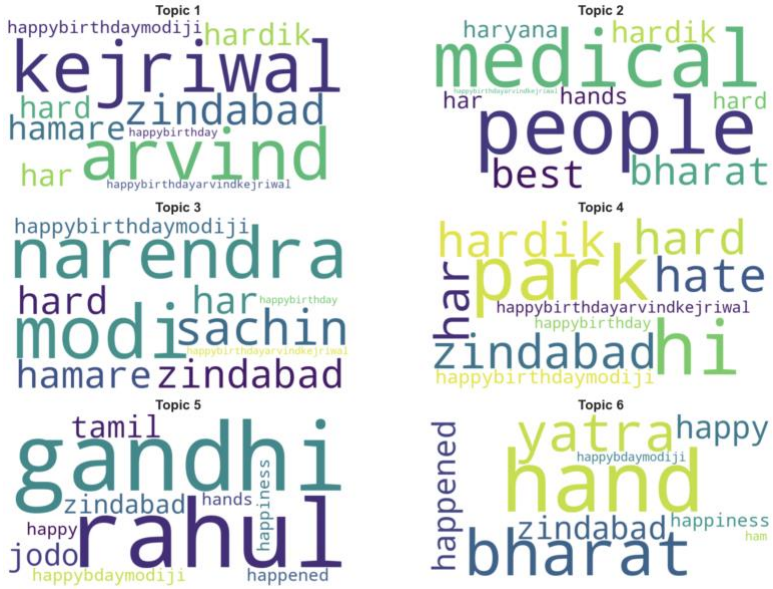


Figure 12: NMF Word Cloud

Comparative study of LSA and NMF Word Cloud:

In this comparison study, we looked at the topics produced by NMF and LSA for a dataset comprising debates about Indian politics, leaders, and events. LSA recognised discussions about political figures, parties, and occasions as well as ambiguity and news. NMF, on the other hand, came up with positive themes, birthday greetings for leaders, and discussions of support and unity within the country. The fact that the two models employed distinct sets of terms to

represent the political, birthday, and national unity themes highlights the subtle variations between the two approaches of extracting topics from text data.

5.2. Topic Labelling:

A crucial step in the analysis of textual data is topic labelling, especially when using Topic Modelling approaches to classify documents into thematic categories. These labels are created using the terms that appear the most frequently used inside each topic, utilising the first three phrases to keep things succinct for all the models. With the help of the generated bar graph, which graphically displays the labelled subjects and their associated tweet counts, we are quickly able to determine how prevalent each theme is in the dataset. The results of complex issue modelling are easier to understand, which also makes it easier to understand and communicate the main findings. It is clear from looking at the subjects generated by each of the four models that some reoccurring themes are shared by all four. Discussions about Rahul Gandhi's ongoing political campaign, Congress elections, and Modi and Arvind's election-related strategies are a few of the recurring topics.

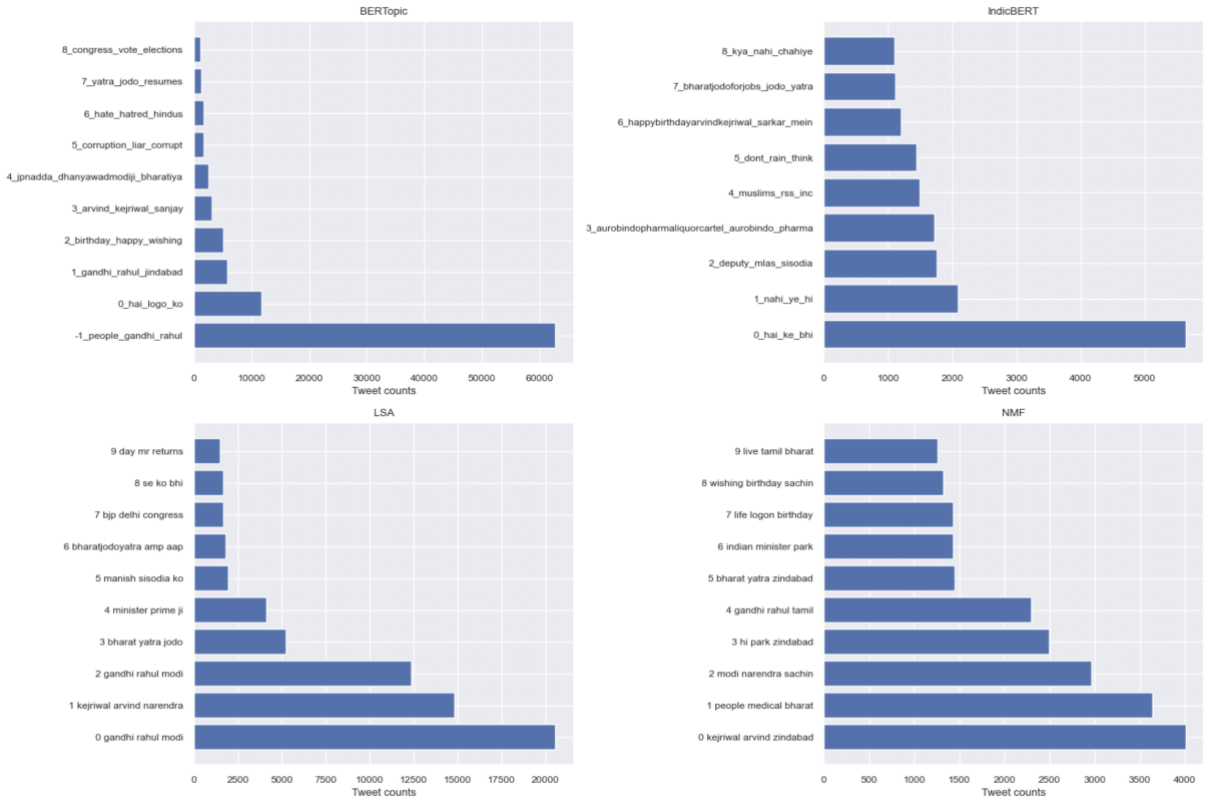


Figure 13: Topic Labelling of Models

5.3. Thematic Analysis:

We have narrowed down the top 10 themes to which we will compare the models in order to get profound insights, by analysing the keywords and personally assigned labels.

5.3.1 Comparison of BERTopic and IndicBERT:

The two models, Bertopic and Indicbert, produce thematic analyses that offer distinctive insights into many facets of Indian politics and political people. Such political issues as Rahul Gandhi's Bharat Jodo Yatra and Modi's Diplomacy are identified and labelled by Bertopic, providing insight into the actual occurrences. Indicbert, on the other hand, concentrates on the feelings, viewpoints, and responses brought on by these events, providing a richer insight of the public and political debate. For instance, while Rahul Gandhi and Narendra Modi are highlighted in Bertopic for their leadership and campaigns, Indicbert focuses on political issues, public opinions, and the nuances of political discourse. Together, these assessments strengthen one another and offer a thorough understanding of the complex realm of Indian politics, from events through public opinion and political responses.

Topic	Keywords - Bertopic	Manual Label - Bertopic	Keywords - Indicbert	Manual Label - Indicbert
1	['gandhi', 'rahul', 'jindabad', 'zindabad', 'bharatjodoyatra', 'jodte', '2024', 'chodo', 'joke', '1monthofbharatjodoyatra']	Rahul Gandhi's Bharat Jodo Yatra	['aurobindopharmaliquorcartel', 'ravivisvesvarayasharadaprasad', 'sisodia', 'manish', 'liquor', 'deputy', 'arrest', 'aurobindopharmascam', 'cbi', 'excise']	Kejriwal Controversies & Politics
2	['birthday', 'happy', 'wishing', 'happybdaymodiji', 'wish', 'happybirthdaymodiji', 'wishes', 'honorable', 'honourable', 'long']	Modi's Birthday Wishes	['nahi', 'ye', 'hai', 'ka', 'hi', 'se', 'ko', 'ho', 'kya', 'gandhi']	Opinions on Rahul Gandhi
3	['arvind', 'kejriwal', 'arvindkejriwal', 'aamaadmiparty', 'cm', 'rajuchadhasharabmafia', 'sisodia', 'best', 'manish', 'rajuchadhawavegroup']	Public Sentiments on Arvind Kejriwal	['dont', 'even', 'know', 'think', 'doesnt', 'would', 'question', 'time', 'hijab', 'rain']	Reaction's to Rahul Gandhi's Yatra
4	['aadmi', 'aam', 'corruption', 'party', 'finished', 'politics', 'mcd', 'corrupt', 'gujarat', 'aap']	Kejriwal's Gujarat Campaign	['ke', 'happybirthdayarvindkejriwal', 'ko', 'hai', 'ne', 'ka', 'aurobindo', 'janmdin', 'bhi', 'mein']	Public Perception of Arvind Kejriwal
5	['kharge', 'mallikarjun', 'congress', 'jindabad', 'jai', 'presidential', 'zindabad', 'president', 'jindabaad', 'vote']	Rahul Gandhi's Leadership	['1monthofbharatjodoyatra', 'womenstandwithbharatjodo', 'bharatjodoinballari', 'women', 'yatra', 'jodo', 'jaishrimahakal', 'bharatjodoyatra', 'completed', 'month']	Women's support for Rahul's Yatra
6	['jpnadda', 'amit', 'shah', 'bjp4ind', 'atal', 'bihari', 'bahadur', 'bjp4up', 'bharatiya', 'jay']	Modi's Leadership Legacy	['bharatjodoyatra', 'rahulgandhizindabad', 'congresszindabad', 'jaicongress', 'yatra', 'karnataka', 'jodo', 'gandhis', 'rahulgandhi', 'sonia']	Rahul's Yatra and political reaction
7	['yatra', 'jodo', 'resumes', 'bharat', 'jindabaad', 'enters', 'avatar', 'via', 'gandhis', 'kerala']	Rahul Gandhi's Yatra	['pharma', 'aurobindo', 'liquorcartel', 'rajuchadhasharabmafia', 'actionagainst', 'scam', 'liquor', 'wave', 'education', 'group']	Arvind Kejriwal and Indian Politics
8	['kerala', 'cow', 'rijil', 'campaigner', 'hindus', 'makutty', 'star', 'bjp', 'slaughtered', 'democracy']	Rahul Gandhi: Controversial Catalyst	['chahiye', 'ye', 'kya', 'ka', 'kejriwal', 'hai', 'arvind', 'bhi', 'nahi', 'se']	Public Sentiments on Arvind
9	['stadium', 'games', '36th', 'opening', 'ceremony', 'ahmedabad', 'sports', 'pmatnationalgames', 'nationalgames2022', 'sardar']	Modi Inaugurates National Games	['rahul', 'even', 'gandhi', 'congress', 'take', 'party', 'think', 'one', 'leader', 'yes']	Perceptions of Gandhi's Leadership in Congress
10	['putin', 'vladimir', 'sco', 'russian', 'samarkand', 'war', 'summit', 'uzbekistan', 'ukraine', 'russia']	Modi's Diplomacy: Ukraine and SCO Summit	['even', 'gandhis', 'like', 'dont', 'pfizer', 'media', 'still', 'deck', 'thats', 'amp']	Political Discourse on Modi and Rahul

Table 1: Thematic Analysis of Top 10 topics for BERTopic and IndicBERT

5.3.2. Comparison of LSA and NMF:

There are significant discrepancies between the ways that NMF and LSA categorise and label topics when compared to the results of manual topic labelling achieved through these methods. With the help of keywords that express feelings and celebrations, NMF-generated labels concentrate on sentiment, joy, and particular events connected to the topics. LSA, in contrast, creates labels that are more logical and politically motivated, highlighting significant individuals and political entities associated with the topics. While NMF's labels appear to be more dispersed and may need additional context for full comprehension, LSA's labels are more explicit and directly tied to Indian political leaders and their activities. The choice between NMF and LSA for topic labelling depends on the particular analytical goals, with LSA offering clearer and more politically-oriented labels and NMF capturing attitudes and celebrations related to the topics.

For Example, let’s consider the two common subjects of NMF and LSA models exhibit, indicating their similar capacity to identify particular themes. First off, talks about the acts and birthday celebrations of Indian Prime Minister Narendra Modi are illustrated by the topics of "Narendra Modi Activities" in NMF and "Modi's Birthday Wishes" in LSA. Second, the topic

titled "Public Sentiments on Arvind Kejriwal" in NMF and "Arvind Kejriwal and AAP" in LSA have a similar resonance, suggesting that both topics have succeeded in encapsulating debates and attitudes regarding Arvind Kejriwal and his political party, AAP. These recurrent subjects highlight the constancy in identifying and characterising important political figures and their associated activities across both NMF and LSA models, despite changes in keyword selection and labelling.

Topic	Keywords	NMF - Manual Labelling	Topic	Keywords	LSA - Manual Labelling
3	kejriwal arvind zindabad hamare hardik hard har happybirthdaymodiji happybirthdayarvindkejriwal happybirthday	Public Sentiments on Arvind Kejriwal	3	kejriwal, arvind, gandhi, rahul, delhi, cm, aap, gujarat, yatra, jodo	Arvind Kejriwal and AAP
16	people medical bharat best hardik haryana hard har hands happybirthdayarvindkejriwal	Perspectives on Indian Leaders	1	modi, narendra, ji, birthday, happy, gandhi, rahul, minister, prime, pm	Modi's Birthday Wishes
2	modi narendra sachin zindabad hamare hard har happybirthdaymodiji happybirthdayarvindkejriwal happybirthday	Narendra Modi Activities	2	gandhi, rahul, narendra, modi, birthday, happy, hai, kejriwal, congress, arvind	Rahul Gandhi and Congress
41	hi park zindabad hate hardik hard har happybirthdaymodiji happybirthdayarvindkejriwal happybirthday	Indian Political Sentiment	4	hai, ki, ke, ko, ka, se, ji, bhi, aur, hi	Opinions on Indian Political Figures
1	gandhi rahul tamil jodo zindabad happybdaymodiji hands happened happiness happy	Perceptions of Rahul Gandhi	5	happy, birthday, modi, narendra, pm, ji, sir, gujarat, bjp, wishing	Modi's Birthday
89	bharat yatra zindabad happy hand hands happened happiness happybdaymodiji ham	Yatra's Political Statement	6	bharat, yatra, jodo, gandhi, rahul, india, gandhis, karnataka, via, pm	Gandhi's Bharat Yatra
35	indian minister park jodo cheetahs media sachin gehlot make ground	Cheetah Conservation Initiative	7	minister, prime, pm, ji, happy, hai, india, chief, birthday, modi	Modi's Leadership
69	life logon birthday opening rain abe happybdaymodiji iss happiness greatest	Leadership and Life: Modi's Birthday	8	ji, happy, birthday, pm, shri, hai, sir, great, bjp, narendramodi	Birthday Wishes for PM Modi
140	wishing birthday sachin minister madhya happiness desh happybirthdaymodiji happybirthdayarvindkejriwal nation	Birthday Wishes for Modi	52	gandhis, free, gandhi, karnataka, rahul, support, dont, country, sonia, education	Support for Rahul and Education
52	live tamil bharat yatra kuno games birthday desh ahmedabad manish	Political Controversies and Statements	43	mr, president, indian, putin, one, next, best, congress, vladimir, war	Indian President and Putin

Table 2: Thematic Analysis of Top 10 topics for LSA and NMF

5.4. Evaluation Metric – Coherence Score:

Bertopic was the best-performing method, with a coherence score of roughly 0.706, demonstrating its ability to produce coherent and semantically significant topics from the dataset. IndicBert came in second place with a score of roughly 0.663, demonstrating its capacity to generate themes with good structure. In comparison, Latent Semantic Analysis (LSA) had a moderate score of roughly 0.508, indicating that topic extraction was relatively coherent but may be improved. Finally, Non-Negative Matrix Factorization (NMF) had the lowest coherence score, at 0.315, indicating difficulties in encoding highly coherent themes. As a result of these findings, BERT-based techniques like Bertopic and IndicBert are now the methods of choice for academics and analysts looking for a deeper comprehension of political issues.

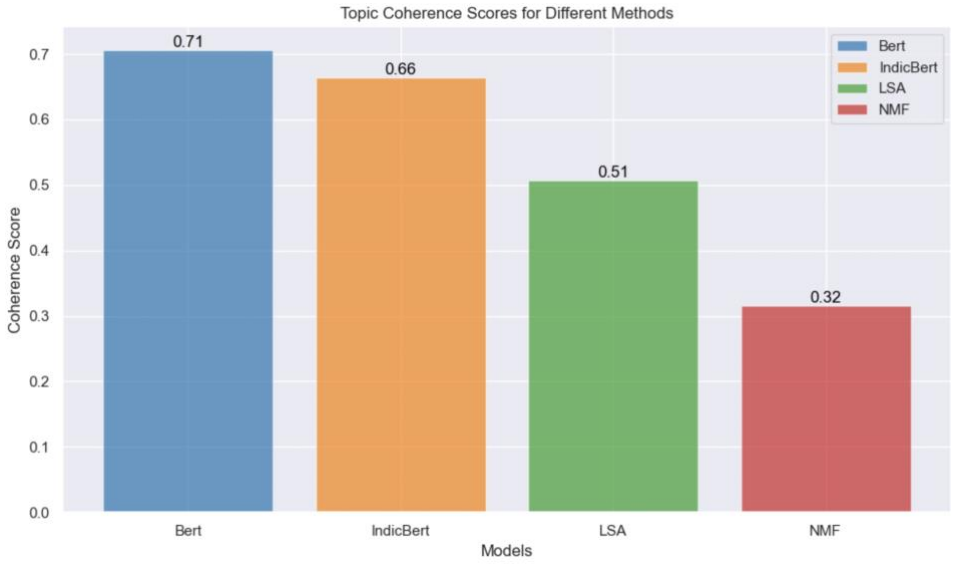


Figure 14: Coherence Score for all four models

6. Limitations and Recommendations for Future Research:

It's important to recognise the constraints and potential directions for further study when using models like IndicBERT, BERTopic, LSA, and NMF. Particularly, the lack of training data and the difficulty of handling code-switching in Indian political writings may restrict IndicBERT's performance. It is crucial to address these problems using domain-specific fine-tuning and code-switching handling approaches. Focused attention is required in the case of BERTopic because to issues with interpretability and scalability, which calls for initiatives to improve the interpretability of generated topics and improve the model's performance for huge political datasets.

Additionally, the shortcomings of LSA, such as its reliance on bag-of-words representations and its inability to capture semantic nuances, highlight the necessity of hybrid models and the creation of contextual LSA variants adapted to Indian political discourse. NMF's sensitivity to initialization and its lack of contextual awareness also point to possible improvements that may be made by utilising robust initialization procedures and contextual information. The success of these models for TM will be greatly improved by addressing these issues. Future studies should keep looking into how well Topic Modelling algorithms work on different platforms and with different algorithms.

Lastly, coherence scores are more suited for structured and well-organized textual material, which presents a drawback when depending only on them to assess Topic Modelling methodologies. There are a number of difficulties and restrictions that must be taken into account when working with unstructured data, such as text from social media or user-generated material. It's critical to broaden the scope of evaluation criteria for TM in future research, considering factors like topic diversity, temporal stability, and human validation. A more comprehensive assessment of TM techniques will be possible because to this wider set of indicators. It is crucial to include human judgement, adapt methodologies to particular domains, and ensure proper data pre-treatment in order to increase the effectiveness of Topic Modelling in complicated circumstances.

7. Conclusion:

To better understand the nuances of Indian politics, we conducted a thorough investigation of Topic Modelling methodologies in the context of Indian political datasets. To have a better understanding of the current political climate, we looked into the Twitter activity of important Indian politicians like Mr. Narendra Modi, Mr. Rahul Gandhi, and Mr. Arvind Kejriwal. By comparing four different Topic Modelling approaches, including LSA, NMF, BERTopic, and IndicBERT, we sought to address important research queries on the efficiency of these techniques in providing meaningful topic labels and cogent themes.

Our investigation produced numerous important conclusions. The best-performing models were BERTopic and IndicBERT, which both had excellent coherence scores (0.71 and 0.66, respectively). This shows that these models produced themes from the Twitter dataset that were consistently coherent and semantically meaningful, making them useful resources for deciphering Indian political conversation. These models' thematic analysis allowed for the discovery of novel political insights in India. While BERTopic concentrated on identifying particular political individuals and events, IndicBERT explored the feelings, opinions, and responses sparked by these events, providing a deeper knowledge of the general public's attitude and political discourse.

LSA and NMF, however, have lower coherence scores despite being still useful. While NMF's labels tended to emphasise sentiment, joy, and unique events, LSA's were more sentimental, political, and logical in nature. Depending on the research goals, one should select either of these topic labelling models: NMF captures attitudes and celebrations associated to themes, while LSA offers more concise, politically oriented labels.

Overall, BERT-based strategies like BERTopic and IndicBERT showed their usefulness in finding and condensing intricate political discussions on Twitter, making them useful tools for political analysts and scholars. But while choosing the most suitable subject modelling approach, it's crucial to take the specific study objectives and the type of data into account.
This study lays the groundwork for further investigation and analysis in this dynamic and important area while also offering insightful analyses of present political discourse in India.

8. References:

1) Meng, Y.; Zhang, Y.; Huang, J.; Zhang, Y.; Han, J. Topic discovery via latent space clustering of pretrained language model representations. In Proceedings of the ACM Web Conference 2022, Lyon, France, 25–29 April 2022; pp. 3143–3152.
2) Blei DM (2012) Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.
3) Dandala, B.; Joopudi, V.; Devarakonda, M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. *Drug Saf.* **2019**, *42*, 135–146.
4) Kastrati, Z.; Arifaj, B.; Lubishtani, A.; Gashi, F.; Nishliu, E. Aspect-Based Opinion Mining of Students' Reviews on Online Courses. In Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, Tianjin, China, 23–26 April 2020; pp. 510–514.
5) Ray, P.; Chakrabarti, A. A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Appl. Comput. Informatics* **2020**, *18*, 163–178.
6) Pennacchiotti, M.; Gurumurthy, S. Investigating topic models for social media user recommendation. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 101–102.
7) Wang, D.; Zhu, S.; Li, T.; Gong, Y. Multi-document summarization using sentence-based topic models. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; World Scientific: Singapore, 2009; pp. 297–300.
8) Tepper, N.; Hashavit, A.; Barnea, M.; Ronen, I.; Leiba, L. Collabot: Personalized group chat summarization. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Los Angeles, CA, USA, 5–9 February 2018; pp. 771–774.
9) Sabeeh, V.; Zohdy, M.; Bashaireh, R.A. Fake News Detection Through Topic Modelling and Optimized Deep Learning with Multi-Domain Knowledge Sources. In *Advances in Data Science and Information Engineering*; Springer: Cham, Switzerland, 2021; pp. 895–907.
10) Wang, T.; Huang, Z.; Gan, C. On mining latent topics from healthcare chat logs. *J. Biomed. Inform.* **2016**, *61*, 247–259.
11) Adanir, G.A. Detecting topics of chat discussions in a computer supported collaborative learning (CSCL) environment. *Turk. Online J. Distance Educ.* **2019**, *20*, 96–114.

12) Agrawal, A.; Fu, W.; Menzies, T. What is wrong with Topic Modelling? And how to fix it using search-based software engineering. *Inf. Softw. Technol.* **2018**, *98*, 74–88.
13) Silveira, R.; Fernandes, C.G.; Neto, J.A.M.; Furtado, V.; Pimentel Filho, J.E. Topic Modelling of legal documents via legal-bert. In Proceedings of the CEUR Workshop, Virtual Event, College Station, TX, USA, 19–20 August 2021; ISSN 1613-0073. Available online: **http://ceur-ws.org** (accessed on 12 October 2022).
14) 210k Tweets on Indian politics- https://www.kaggle.com/datasets/soumendraprasad/201k-tweets-on-mrmodimrrahulmrkejrielecanal
15) Gabarron E, Dorronzoro E, Reichenpfader D, Denecke K. What Do Autistic People Discuss on Twitter? An Approach Using BERTopic Modelling. Stud Health Technol Inform. 2023 May 18;302:403-407. doi: 10.3233/SHTI230161. PMID: 37203705.
16) Athukorala, S., Mohotti, W. An effective short-text Topic Modelling with neighbourhood assistance-driven NMF in Twitter. *Soc. Netw. Anal. Min.* **12**, 89 (2022). https://doi.org/10.1007/s13278-022-00898-5
17) H. Gupta and M. Patel, "Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, 2021, pp. 511-517, doi: 10.1109/ICAIS50930.2021.9395976.
18) IndicBERT based approach for Sentiment Analysis on Code-Mixed Tamil Tweets R.Ramesh Kannan, Ratnavel Rajalakshmi and Lokesh Kumar. https://ceur-ws.org/Vol-3159/T3-16.pdf
19) Discovery," in 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018), 2019.
20) C.-K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with Topic Modelling," Scientometrics, vol. 100, no. 3, pp. 767–786, 2014.
21) H. Jelodar et al., "Latent Dirichlet Allocation (LDA) and Topic Modelling: models, applications, a survey," Multimed. Tools Appl., vol. 78, no. 11, pp. 15169–15211, 2019.
22) K. Hornik and B. Grün, "topicmodels: An R package for fitting topic models," J. Stat. Softw., vol. 40, no. 13, pp. 1–30, 2011.
23) Z. Tong and H. Zhang, "AText MINING RESEARCH BASED ON LDA TOPIC MODELLING," Int. Conf. Comput. Sci. Eng. Inf. Technol., pp. 201–210, 2016.
24) Grootendorst, M. (2022). BERTopic: Neural Topic Modelling With a Class-Based TF-IDFProcedure. arXiv:2203.05794v0571.Availableonline at: https://arxiv.org/pdf/2203.05794.pdf (accessed March 15, 2022).
25) Martengr, https://maartengr.github.io/BERTopic/algorithm/algorithm.html
26) Huggingface, https://huggingface.co/ai4bharat/indic-bert
27) Deerwester et al., *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, Volume 41, Issue 6 p. 391–407, 1990 (link).
28) Vavasis Stephen A.. 2010. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization* 20, 3 (2010), 1364–1377
29) Emil Rijcken, Cv Topic Coherence Score, https://towardsdatascience.com/c_v-topic-coherence-explained-fc70e2a85227
30) Queiroz, M. M. (2018). A framework based on Twitter and big data analytics to enhance sustainability performance. *Environ. Qual. Manag*. 28, 95–100. doi: 10.1002/tqem.21576
31) Femenia-Serra, F., Gretzel, U., and Alzua-Sorzabal, A. (2022). Instagram travel influencers in #quarantine: communicative practices and roles during COVID-19. *Tour. Manag*. 89:104454. doi: 10.1016/j.tourman.2021.104454

32) Egger, R. (2022b). “Topic Modelling. Modelling hidden semantic structures in textual data,” in *Applied Data Science in Tourism. Interdisciplinary Approaches, Methodologies and Applications*, ed R. Egger (Berlin: Springer), 18. doi: 10.1007/978-3-030-88389-8_18
33) Obadimu, A., Mead, E., and Agarwal, N. (2019). “Identifying latent toxic features on YouTube using non-negative matrix factorization,” in *The Ninth International Conference on social media Technologies, Communication, and Informatics* (Valencia), 1–6.
34) Albalawi, R.; Yeap, T.H.; Benyoucef, M. Using Topic Modelling Methods for Short-Text Data: A Comparative Analysis. *Front. Artif. Intell.* **2020**, *3*, 42
35) Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 2–6 February 2015; pp. 399–408.