INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Textual Data Analysis of the Union Budget of India

Kirti Makwana[1]*

[1] Assistant Professor, Faculty of Management Studies (FMS), Indukaka Ipcowala Institute of Management (I2IM), Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India

OPEN ACCESS

Received: 15-05-2023

Accepted: 23-12-2023

Published: 31-01-2024

Citation: Makwana K (2024) A Textual Data Analysis of the Union Budget of India. Indian Journal of Science and Technology 17(5): 478-486. https://doi.org/10.17485/IJST/v17i5.1174

* **Corresponding author**.

kirtimakwana.mba@charusat.ac.in, kirtimakwana85@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To present a textual data analysis of the Union Budgets of India for financial years 2019-20 to 2023-24). Examining the policy narratives, key announcements, and thematic emphasis in Budget speeches to explore about the government's priorities. **Methods:** The analysis is centered on the budget presented by Nirmala Sitharaman, the Finance Minister of India. The study emphases on the budget speeches conveyed by her and serves a combination of quantitative and qualitative techniques to classify the main themes and primacy of the budget. To categorize the discourse into diverse subjects, Natural Language Processing (NLP), Topic Modeling and Sentiment Analysis methods are used. **Findings:** The outcomes recommend that the budget emphases on encouraging commercial growth, improving living standard, and providing liberation to many sectors affected by the COVID-19 pandemic. "India", "Government", "Infrastructure", "Sector" etc. are some of the words which are used repeatedly in each budget presented. Further high value of Term Frequency-Inverse Document Frequency (TF-IDF) suggests that India (0.323), Government (0.315), Tax (0.268), Crores (0.380) are some of the words which are the most important and relevant words during Budget presentations. Correlation matrix suggests that topic 1 is highly negatively correlated with topic 2 (coefficient value – (-0.832)). The paper concludes by deliberating the repercussions of the budget on the Indian economy and the challenges that are to be addressed to attain the budget's intents. **Novelty :** Largely, the research paper delivers an all-inclusive understanding of the Indian Union Budget and its possible influence on the country's economic and social development.

Keywords: Textual Data Analysis; Union Budget; Nirmala Sitharaman; Bag of Words; Sentiment Analysis; TF-IDF

1 Introduction

The Indian government's fiscal objectives for the financial year are defined in the Union Budget, a crucial document. The nation's yearly budget is known as the Indian Union Budget, usually denoted to as the annual Financial Report (Article 112 of the Indian Constitution). To understand the government's intentions and plans, economists, legislators, and specialists scrutinize the budget in advance it is presented to Parliament by India's finance minister. Since 1947 (after independence) a total of 73 yearly, 14 interim and 4 special budgets/mini budgets have been presented. Nirmala Sitharaman, the Indian Finance Minister, presented her fifth successive budget for the fiscal year 2023–2024. As of 1 February 2019 for Fiscal Year 2019–20, 2020–21, 2021–22, 2022– 23, and F.Y. 2023-24, respectively,

Textual Data Analysis convey the procedure to extract important discernment and details from textual data. It is an essential element in the data analytics field. It is used to analyse enormous volumes of text data, such as customer sentiments, social media columns, and press release articles. It gives valued insights into several themes and trends. With the propagation of digital media, textual data analysis has turn out to be a critical instrument for industries to make informed resolutions, improve customer satisfaction, and gain a competitive gain.

Textual Data Analysis of Union budget scrutinizes the policy narratives, highlights important declarations, and differentiate thematic emphasis within Budget speeches, targeting to gain insights into the government's priorities. This research paper includes use of NLP methods to clean and pre-process the data, followed by several statistical and machine learning algorithms to recognize patterns, themes, and sentiment in the text. In essence, the study provides to the requirements of making well-informed pronouncements, encouraging transparency in governance, and adopting a more profound comprehension of the government's economic priorities throughout the designated financial years.

1.1 Traditional Framework for Text Analysis

In the subsequent section, the brief indication of the essential structure of text analytics is presented, which is used to analyse a text corpus. Usually, a traditional text analytics outline contains four main stages. These stages are illustrated in Figure 1.

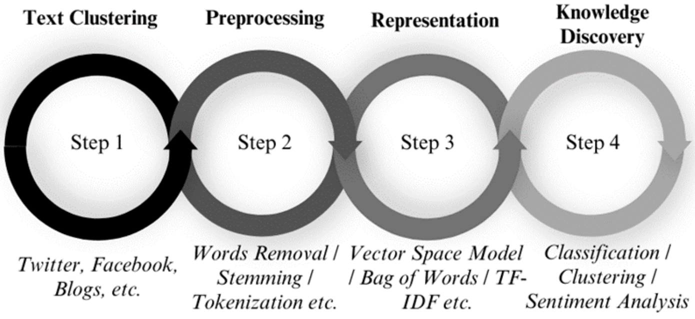

Text Clustering
Preprocessing
Representation
Knowledge Discovery
Step 1
Step 2
Step 3
Step 4
Twitter, Facebook, Blogs, etc.
Words Removal / Stemming / Tokenization etc.
Vector Space Model / Bag of Words / TF-IDF etc.
Classification / Clustering / Sentiment Analysis

Fig 1. Traditional Framework for Text Analysis (Source: Created by the Author)

The existing research work available on textual data analysis of the Indian Union budget underlines the significance of using data-driven methods to analyse the government's financial strategies. Sentiment analysis, topic modelling, and network analysis can offer important understandings into the government's priorities and strategies and notify policy discussions and deliberations. Further research in this area can help deepen our understanding of the Union Budget of India and its implications for the economy and society.

Several studies have examined the Union Budget of India using textual data analysis techniques such as sentiment analysis, topic modelling, and network analysis. These studies have analysed the budget documents to identify the dominant themes, sentiment, and key policy measures proposed by the government.

Tweets on the 2016 Indian Union budget were analysed for sentiment[1]. In this study, a lexicon-based sentiment analysis was used along with a dictionary of positive and negative words. Using the dataset, the investigators examined word co-occurrences. Tax-related terms were the most frequent, which reflects public opinion. Youth and rural populations contributed little to Twitter conversations. Textual data analysis has recently become a popular method for sorting extensive and vast textual content. As a rule, it is a matter of extracting hidden themes and explaining the documents on the basis of these themes.[2]

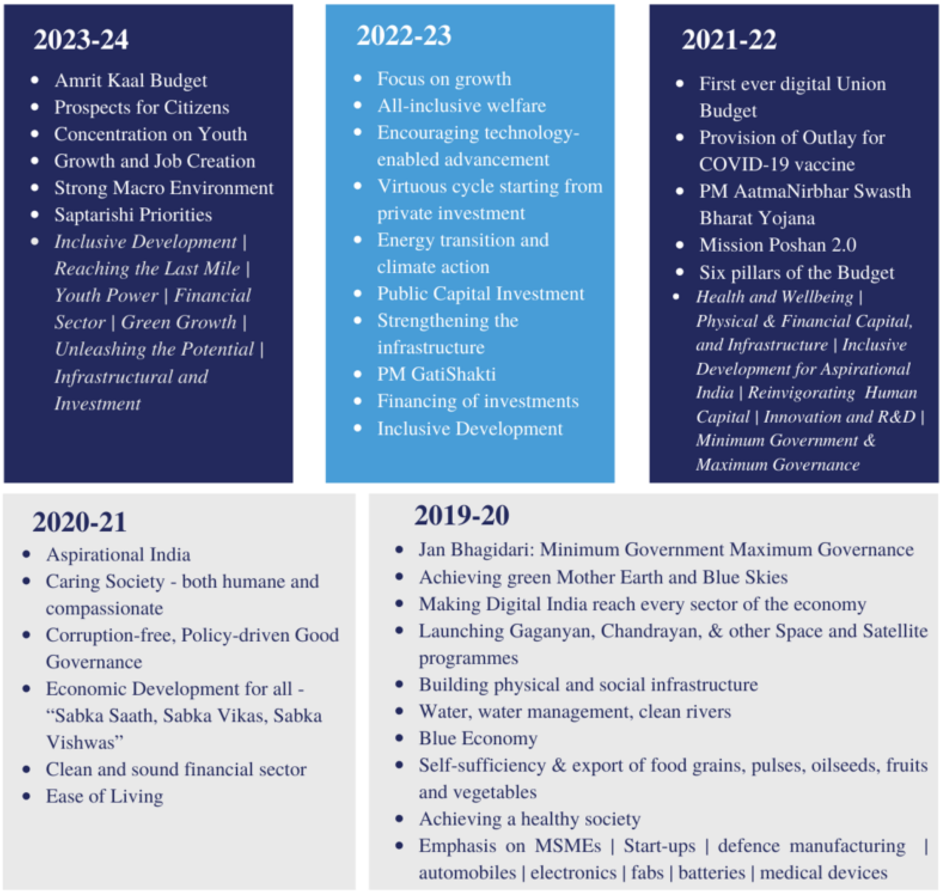
The literature shows that different methods and procedures that can be used to excerpt unseen info after gathering of documents. Some of the research work have investigated with related concepts and similarly pointed out that there are diverse illustrations of theme models, and they use this approach remarkably.[3]

Various extensions to the primary theme model have been developed, such as theme models for both text and images[4] author topic models[5], author role topic models[6],and hidden Markov topic models, which can distinguish between semantic and syntactic topics[7].

2 Methodology

2.1 Data Collection

The researchers collected English versions of Union Budget Speeches presented by Nirmala Sitharaman, Minister of Finance, India, accessible on https://www.indiabudget.gov.in/. February 2023, Nirmala Sitharaman delivered her fifth sequential Union Budget presentation. The researchers considered all five consecutive Budget presentations (Financial Year 2019-20 to Financial Year 2023-24). Further, Python programming and its packages on Orange Data Mining Version 3.34.0[(8–10)] are used to carry out textual data analysis. The summary/ highlights of data collection is described in the following figure:



2023-24

- Amrit Kaal Budget
- Prospects for Citizens
- Concentration on Youth
- Growth and Job Creation
- Strong Macro Environment
- Saptarishi Priorities
- *Inclusive Development | Reaching the Last Mile | Youth Power | Financial Sector | Green Growth | Unleashing the Potential | Infrastructural and Investment*

2022-23

- Focus on growth
- All-inclusive welfare
- Encouraging technology-enabled advancement
- Virtuous cycle starting from private investment
- Energy transition and climate action
- Public Capital Investment
- Strengthening the infrastructure
- PM GatiShakti
- Financing of investments
- Inclusive Development

2021-22

- First ever digital Union Budget
- Provision of Outlay for COVID-19 vaccine
- PM AatmaNirbhar Swasth Bharat Yojana
- Mission Poshan 2.0
- Six pillars of the Budget
- *Health and Wellbeing | Physical & Financial Capital, and Infrastructure | Inclusive Development for Aspirational India | Reinvigorating Human Capital | Innovation and R&D | Minimum Government & Maximum Governance*

2020-21

- Aspirational India
- Caring Society - both humane and compassionate
- Corruption-free, Policy-driven Good Governance
- Economic Development for all - "Sabka Saath, Sabka Vikas, Sabka Vishwas"
- Clean and sound financial sector
- Ease of Living

2019-20

- Jan Bhagidari: Minimum Government Maximum Governance
- Achieving green Mother Earth and Blue Skies
- Making Digital India reach every sector of the economy
- Launching Gaganyan, Chandrayan, & other Space and Satellite programmes
- Building physical and social infrastructure
- Water, water management, clean rivers
- Blue Economy
- Self-sufficiency & export of food grains, pulses, oilseeds, fruits and vegetables
- Achieving a healthy society
- Emphasis on MSMEs | Start-ups | defence manufacturing | automobiles | electronics | fabs | batteries | medical devices

Fig 2. Key Highlights of Union Budgets (2019-20 Onwards) (**Source:** Created by the Author)

2.2 Pre-processing

In any unstructured textual data analysis task, cleaning or pre-processing of the data is an important step. Some of the text pre-processing stages are – removal of punctuation, stop wards, emoji, emoticons, URL etc., stemming, lowering case, stemming, lemmatization, conversion of emoticons and emoji to words, and spelling corrections (if any).

2.3 General Framework

The Textual Data Analytics General Framework, powered by Orange Software, presents a well-organized and adaptable method for extracting valuable insights from textual data (Figure 3). Orange Software acts as the core tool for implementing this framework, providing a user-friendly and interactive platform for data analysis. The essential elements of this framework encompass:

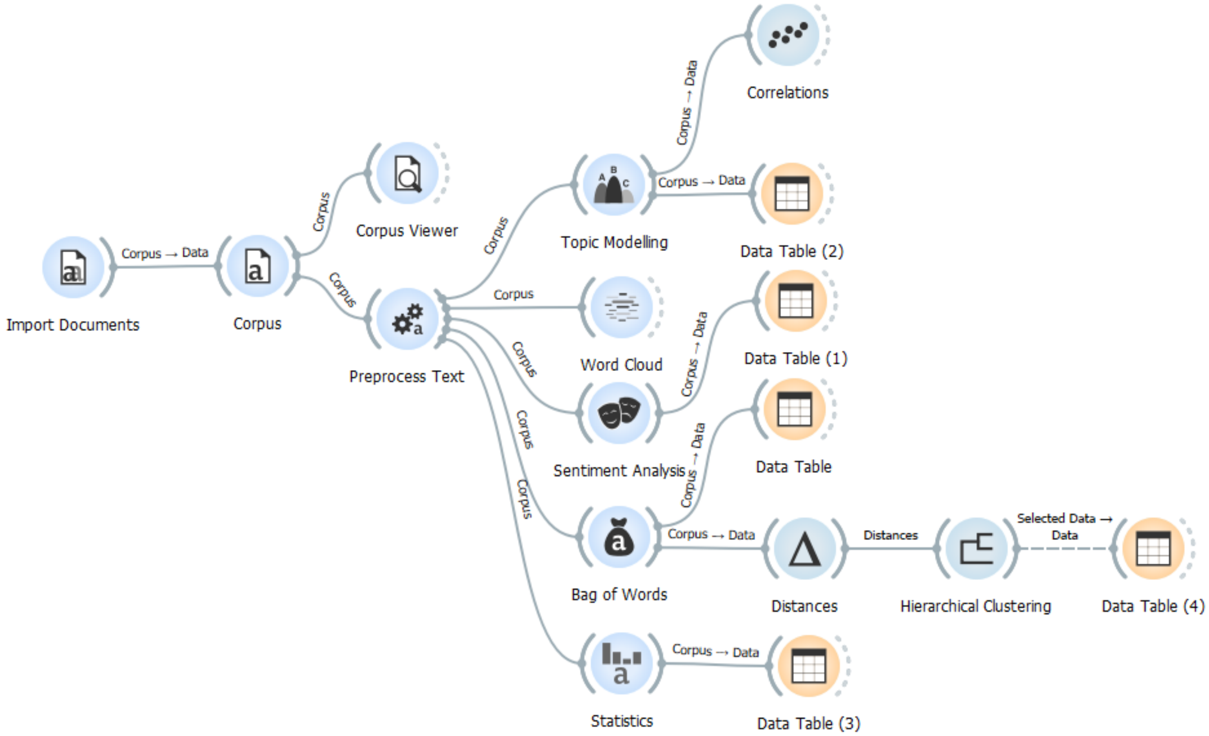


Fig 3. General Framework of Textual Data Analytics Using Orange Software (Source: Created by the Author)

- **Data Preparation** - Textual data is imported into the Orange Software as the first step of the framework. Cleaning and pre-processing are then carried out on the textual data.
- **Text Mining Techniques** - Tokenization is a technique utilized by Orange Software to dissect textual data into significant units, such as words or phrases. The framework employed by Orange Software extracts pertinent features from the text, transforming it into a format that is appropriate for analysis. This process may encompass methods like TF-IDF (Term Frequency-Inverse Document Frequency) or word embedding's.
- **Exploratory Data Analysis (EDA)** - EDA commonly utilizes word clouds, frequency distributions, and topic modelling visualizations to facilitate analysis and understanding.
- **Text Analytics Algorithms** - The sentiment analysis tools incorporated in the framework enable the evaluation of the sentiment conveyed in the text, thereby offering valuable insights into the emotional essence of the content. Topic Modelling facilitates algorithms such as Latent Dirichlet Allocation (LDA) to detect dominant themes or subjects present in the textual data.
- **Machine Learning Integration** - Incorporates machine learning algorithms to perform tasks such as text classification and clustering. This enables the automated categorization of text or the grouping of similar documents.
- **Export and Reporting** - Exports the analysed results in various formats for further reporting or integration with other tools.

2.3 Objectives

1. To conduct a comprehensive textual data analysis of the Union Budgets of India presented by Nirmala Sitharaman, the Finance Minister of India.
2. To identify the key themes and priorities of the Union Budgets.
3. To analyse the overall sentiment and tone of the Union Budget of India.
4. To develop a framework for textual data analysis of budgetary documents and apply it to the Union Budget of India to generate insights and recommendations for future research and policy-making.

3 Results and Discussion

Table 1. Top Ten Words Spoken (Frequency) in Union Budgets' Presentation

2019-20		2020-21		2021-22		2022-23		2023-24	
Word	**F**	**Word**	**F**	**Word**	**F**	**Word**	**F**	**Word**	**F**
India	80	Would	80	Crores	97	Government	37	Crore	60
Government	77	Tax	78	Year/ Years	61	Digital	34	Lakh	51
Year/Years	54	Government	77	Propose	50	Capital	33	Tax	43
Tax	43	India	60	Tax	46	Crore	31	Development	28
Crore	36	Crore	60	Government	41	Development	29	Government	28
Scheme	35	Propose	55	India	40	Provide	27	Infrastructure	26
Propose	33	Lakh	54	Infrastructure	37	Tax	26	India	24
Would	32	Shall	41	Lakh	34	Infrastructure	25	Income	24
Lakh	28	New	37	Budget	32	Investment	25	Financial	24
Investment	27	Income	36	Scheme	26	States	21	Digital	21

The ”Most Frequent Used Words in Union Budget”, refers to the words or phrases that appear most often in the document. This can give insights into the key priorities and focus areas of the government, as well as the broader trends and themes in the budget. To define the utmost recurrent used words in the Budget speech, the investigator has applied text mining method. This includes analysing the transcript of the budget to find the occurrence of individual word or phrase. Table 1 illustrates the most recurrent words in union budgets' speech presented by Nirmala Sitharaman, the Finance Minister of India. As perceived, “India”, “Government”, “Infrastructure”, “Sector” etc. are some of the words which are used recurrently in each budget presented. Analysing the most recurrent words during the Union Budget presentation, can give understandings into the government's primacy, purposes, attention areas, and policy guidelines. For example, if the word ”infrastructure” appears often, it proposes that the government is placing a lot of stress on evolving infrastructural facilities in the country.

In the age of big data and information excess, mining the utmost pertinent and important information from documentary data has developed a critical job. One such assignment is Key Word Extraction, which includes repeatedly classifying and extracting the utmost applicable words or expressions that signify the key subjects and themes in a specified text or corpus. Natural Language Processing (NLP) methods have demonstrated to be extremely effective in Key Word Extraction by leveraging numerous language topographies and statistical approaches to classify and excerpt significant relations and expressions from a text corpus.

Term Frequency-Inverse Document Frequency (TF-IDF) is a common and extensively used techniques for keyword extraction in NLP. TF-IDF score represents the significance of the term in the text and the corpus as a whole. By calculating the TF-IDF value for each term in a document or corpus, we can identify the most important and relevant terms, which are often used as keywords or topics for further analysis. Table 2 shows the TF-IDF values of the Year wise budgets.

A high TF-IDF score conveys that the term/word is both important in the document and relatively rare in the corpus, suggesting that it is a significant keyword or topic.

3.1 Topic Correlation Analysis

- **Pearson Correlation**

As shown in the correlation matrix of Table 3, five topics generated are taken into consideration. It is found that, topic 1 is highly negatively correlated with topic 2 (coefficient value – (-0.832)) which says that, topic 1 and topic 2 travel in opposite directions.

- **Word Clouds of Union Budget Speeches**

A word cloud is a graphic demonstration of the occurrence of words in a text. It's a cluster of words with the most frequent words seem larger than other words. A word-frequency exploration of Finance Minister Nirmala Sitharaman's budget speeches suggests the Finance Minister used the words in her speeches were supported deeply by intents and government's strategies. The figures below show word clouds of Union Budgets for financial years 2019-20, 2020-21, 2021-22, 2022-23, 2023-24 and comprehensive (all five financial years). Word clouds show the themes that got the most prominence in the budget. The size of the word in the image corresponds to their frequency in the budget speech. As seen “infrastructure”, “digital”, “tax”, “development”,

Table 2. TF-IDF Values (Year wise Budgets)

2019-20		2020-21		2021-22		2022-23		2023-24	
Word	TF-IDF	Word	TF-IDF	Word	TF-IDF	Word	TF-IDF	Word	TF-IDF
India	0.323	Tax	0.268	Crores	0.380	Government	0.176	Crore	0.279
Government	0.315	Would	0.262	Propose	0.196	Digital	0.162	Lakh	0.237
Tax	0.186	Government	0.252	Tax	0.180	Capital	0.157	Tax	0.200
Crore	0.145	India	0.196	Government	0.164	Crore	0.148	Government	0.134
Scheme	0.141	Crore	0.196	India	0.160	Development	0.138	Development	0.130
Propose	0.133	Propose	0.180	Infrastructure	0.145	Provide	0.129	Infrastructure	0.120
Would	0.129	Lakh	0.177	Lakh	0.133	Tax	0.133	India	0.116
Lakh	0.113	Shall	0.134	Budget	0.125	Infrastructure	0.124	Income	0.116
Investment	0.113	Income	0.131	Would	0.102	Investment	0.124	Financial	0.112
Year/Years	0.109	New	0.121	Scheme	0.102	Lakh	0.109	Digital	0.098
Infrastructure	0.101	Year/Years	0.121	Capital	0.098	Income	0.105	New	0.098
Sector	0.101	Rate	0.108	National	0.098	States	0.100	Green	0.098
Public	0.097	Sector	0.101	Health	0.094	Customs	0.095	Sector	0.093
Financial	0.097	Provide	0.101	Sector	0.094	India	0.090	Duty	0.093

Table 3. Correlation Matrix

Topic No.	1	2	3	4	5
1	1	-0.832	+0.384	+0.377	-0.049
2		1	+0.011	+0.011	-0.001
3			1	-0.005	0.001
4				1	0.001
5					1

“government”, “investment” etc. are most recurring words in Nirmala Sitaraman’s speeches. The words, ’development’ and ’infrastructure’, reflect the way ruling government intends to enhance development and growth by massively rising outlay on infrastructure building such as highways and railways. The less frequent words mentioned in the word clouds represents that, these themes have a lower priority in the budget (Figure 4).

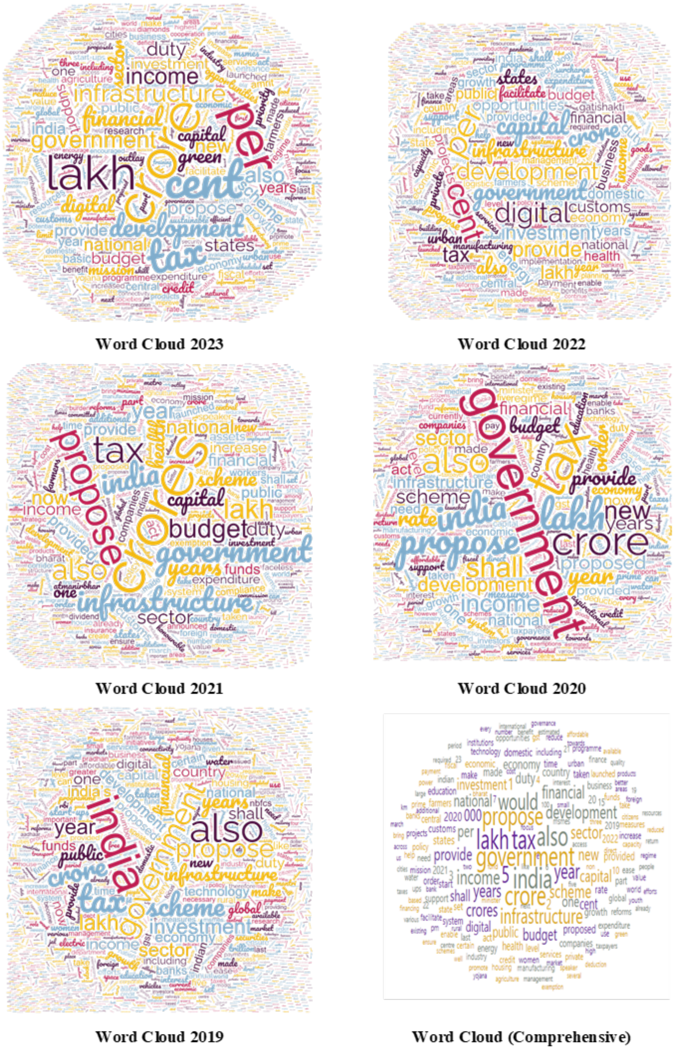
- **Sentiment Analysis**

Sentiment analysis, also denoted as ‘opinion mining’, is a method to Natural Language Processing (NLP) which classifies the emotive tone of a body of text. Sentiment analysis involves recognizing and classifying the emotions articulated in a given transcript, whether positive, negative, or neutral. Sentiment analysis is carried out to determine if the union budget is positive, negative or neutral. To identify the sentiments of the union budgets VADER (Valence Aware Dictionary for Sentiment Reasoning) is used. Predominantly, VADER sentiment analysis depends on a vocabulary that maps verbal features to sentiment intensities/sentiment scores. This method gives sensitivity to both polarity (positive/negative) and intensity (strength) of emotion. The table below shows sensitivity of last five years’ budgets:

Table 4. Sentiment Analysis of Union Budgets

Sr. No.	Financial Year	Positive Sentiment	Negative Sentiment	Neutral Sentiment
1	2019-20 Union Budget	0.147	0.022	0.831
2	2020-21 Union Budget	0.136	0.023	0.840
3	2021-22 Union Budget	0.138	0.032	0.830
4	2022-23 Union Budget	0.151	0.021	0.827
5	2023-24 Union Budget	0.154	0.019	0.826

It is observed from the Table 4 that a majority of words used by the Finance Minister of India during presentation of Union Budgets (Financial Year 2019-20 onward) fell into the neutral sentiments with minor percentage of negative and positive



Word Cloud 2023

Word Cloud 2022

Word Cloud 2021

Word Cloud 2020

Word Cloud 2019

Word Cloud (Comprehensive)

Fig 4. Word Clouds of Budget Speeches (Source: Created by the Author)

sentiment categories.

- **Heat Map**

Heat maps are inordinate to visualize analogous numeric variables. The colour scheme used in a heat map is designed to make it easy to interpret and understand the data being analysed. By using bright colours to represent positive compound values and dull colours to represent negative compound values, the heat map provides a clear visual representation of the data, allowing patterns and trends to be easily identified. The color-coding scheme used in the analysis represents the sentiment of the documents. Positive sentiment documents are denoted by white-coloured bars with positive compound values, whereas negative sentiment documents are depicted by blue-coloured bars with negative compound values. The same is seen in the Figure 5.

- **Hierarchical Clustering**

In data mining, hierarchical clustering is a clustering method that generates a hierarchical configuration of clusters in a dataset created on the connection or distinction amongst data points. This technique is acknowledged as hierarchical cluster analysis and delivers a graphical illustration of the association among the clusters, which can support data exploration and understanding. Primarily, individually data point is considered as a separate cluster. Formerly, the algorithm increasingly

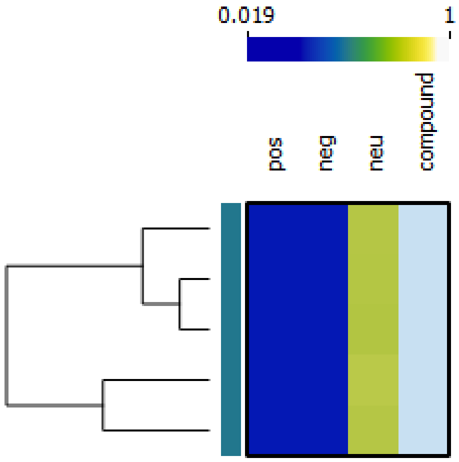


Fig 5. Heat Map of Union Budgets (2019-20 onward) (Source: Created by the Author)

combines the neighbouring clusters, lasting this procedure till a predetermined stopping condition is seen. A tree-shaped diagram, known as a dendrogram is created as an outcome of hierarchical clustering, representing the hierarchical associations amongst the clusters. It offers a graphical depiction of in what way related or divergent the clusters are to each other and permits for the identification of sub clusters or outlier data points in the dataset. The Figure 6 shows hierarchical clustering of the union budget speeches. As seen in the Figure 6, main four clusters are made.

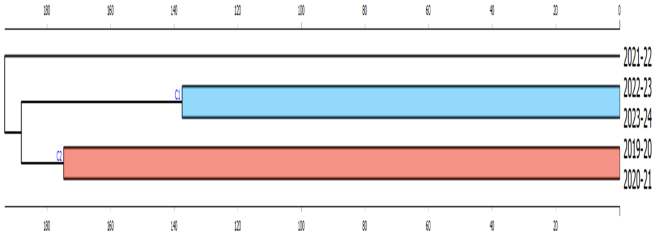


Fig 6. Hierarchical Clustering (Source: Created by the Author)

4 Conclusion

This textual data analysis of the Union Budget of India presented by Finance Minister Nirmala Sitharaman for the years 2019-20 to 2023-24 provides insight into the major themes, priorities, and sentiments of the budget speeches. To extract valuable insights, Natural Language Processing (NLP), Topic Modeling, and Sentiment Analysis were used. Several noteworthy aspects are revealed by this analysis. It appears that the government's priority is to support economic growth, improve living standards, and tackle COVID-19's challenges. The budget presentations consistently included key terms such as "India," "Government," "Infrastructure," "Tax," and "Crores," emphasizing their importance in government policy discussions. In each budget presentation, the Term Frequency-Inverse Document Frequency (TF-IDF) analysis revealed specific words that were relevant and important. High TF-IDF scores for terms like "India," "Government," "Tax," and "Crore" underscored their critical roles in shaping budget priorities. Correlation matrix analysis revealed interesting relationships between the various budget topics, with some topics displaying strong negative correlations, suggesting divergent emphasis directions. With VADER, sentiment analysis showed a majority of budget content to be neutral, and a small percentage to be positive or negative. The budget speeches were clustered hierarchically, illustrating how related or divergent certain topics and themes were. Overall, this textual data analysis has provided valuable insights into the Union Budget of India and its potential economic and social implications. Utilizing advanced analytical techniques, this research aims to enhance understanding of the government's priorities and strategies, contributing to future policy discussions. By using data-driven approaches, we can better understand the intricate nature of budgetary documents and the impact they have on a nation's trajectory.

References

1) Shakeel M, Karwal V. Lexicon-based sentiment analysis of Indian Union Budget 2016–17. In: 2016 International Conference on Signal Processing and Communication (ICSC). IEEE. 2017;p. 299–302. Available from: https://doi.org/10.1109/ICSPCom.2016.7980595.
2) Makwana K, Ganatra AP. Textual Data Analysis of 'Mann Ki Baat' Show. *Indian Journal Of Science And Technology*. 2022;15(37):1859–1867. Available from: https://doi.org/10.17485/IJST/v15i37.848.
3) Dredze M, Wallach HM, Puller D, Pereira F. Generating summary keywords for emails using topics. In: and others, editor. Proceedings of the 13th international conference on Intelligent user interfaces. ACM. 2008;p. 199–206. Available from: https://doi.org/10.1145/1378773.1378800.
4) Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 2003;3:993–1022. Available from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.
5) Rosen-Zvi M, Chemudugunta C, Griffiths T, Smyth P, Steyvers M. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*. 2010;28(1):1–38. Available from: https://cocosci.princeton.edu/tom/papers/AT_tois.pdf.
6) Mccallum A, Corrada-Emmanuel A, Wang X. Topic and role discovery in social networks. In: IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence. 2005;p. 786–791. Available from: https://dl.acm.org/doi/10.5555/1642293.1642419.
7) Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB. Integrating topics and syntax. In: NIPS'04: Proceedings of the 17th International Conference on Neural Information Processing Systems. 2004;p. 537–544. Available from: https://dl.acm.org/doi/10.5555/2976040.2976108.
8) Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, et al. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*. 2013;14:2349–2353. Available from: https://www.jmlr.org/papers/volume14/demsar13a/demsar13a.pdf.
9) Rajiv SSC, Singh AK. Defence Budget 2023-24: Trend Analysis. 2023. Available from: https://doi.org/10.13140/RG.2.2.35712.10241.
10) Kaushal N, Ghalawat S, Saroha A. Communicating Five-Year Budgets for the Indian Economy: Comparative Text and Sentiment Analysis. *Journal of Content, Community & Communication*. 2021;14(7):133–144. Available from: https://doi.org/10.31620/JCCC.12.21/11.