Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

School of Social Sciences and Philosophy
Assignment Submission Form

Student Name:	Shekhar Kedia
Student ID Number:	23351315
Programme Title:	MSc Applied Social Data Science
Module Title:	Quantitative Text Analysis for Social Scientists
Assessment Title:	Research Project
Lecturer (s):	Dr. Martyn Egan
Date Submitted:	8th Mar '24

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar

I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write

Shekhar kedia

Signed: ______________________________

Date: **8th Mar '24**

Table of contents

Table of figures

India Budget Analysis:
Text complexity and net sentiment

1 Background

India ended 2023 on a high note reporting robust economic growth. The statistics office reported Gross domestic product (GDP) growth of 8.4% in the October to December quarter. This is positive news for the current Prime Minister, Narendra Modi as India is set for a general election due in April and May in 2024, potentially marking his third term if successful. The International Monetary Fund projects India's economy to grow by 6.7% for the fiscal year through March and 6.5% in the fiscal year starting April (World Economic Outlook, 2024). The surge in strong GDP numbers has also led to high sentiment among people and the Indian stock markets are at all time high. Additionally, numerous multinational firms are expressing interest in establishing operations within the thriving Indian market.

It's essential to recognize that this economic surge is not an isolated event but rather the result of consistent efforts spanning the past 77 years. Independent India is still a young and growing economy. For such a young and growing economy, the union budget then becomes the fulcrum of development. Every year on February 1st, the presiding Finance Minister presents the Union Budget in the Parliament which essentially reflects the roadmap of the fiscal plan of the government and the details of allocations to various sectors. The Indian economy primarily consists of three key sectors viz. primary (agricultural), secondary (manufacturing and industry), and tertiary (others).

Being a democratic nation, political parties having different agenda and ideological standpoints have been in power. The Indian National Congress (INC) was the first party to be in power and ruled straight for more than three decades which is considered to be left-winged in the political spectrum. Since 2014, India is governed by the Bhartiya Janta Party (BJP), led by Narendra Modi which is considered to be right-winged in its perspective. However, it is important to note that economic reforms have always been on the forefront of any government.

This study aims to understand the key agenda points or the key words that have been highlighted in the Union budget. It also aims to understand how the lexical complexity of the Union budget have changed over time and if the INC government used more complex language in the Union Budget. Lastly, the study also looks at how the net sentiment of the budget has changed over the past 77 years.

1.1 Contribution

The Union budget should be easy for Indian general mass to understand. The understanding of language complexity and readability of the budget speeches thus becomes crucial. Moreover, the key words reflected in the budget speeches of the finance ministers provide an indication of priorities of the government which then is becomes crucial to understand.

The estimation of overall net sentiment value of the speeches and how it has changed over the time will help government officials and policymakers understand the pattern and guide future speech narratives. Researchers can also make use of the findings to corroborate with their research and also correlate with net sentiment value of general public.

2 Related works

The current state of literature on Indian Union Budget Analysis reveals a notable gap, particularly in the examination of sentiment within budget speeches. While existing research highlights public sentiment around the budget, there is a limited exploration of sentiment within the speeches themselves. Moreover, the literature is silent on the assessment of the lexical complexity of language used in the budget speeches and an understanding of its readability. However, studies have shown that government use complex political texts these days to address the general population to ensure a wider acceptance (Bischof & Senninger, 2017). There is also evidence that suggests political texts are getting shorter and less complex(Benoit, 2019).

The recent study which conducted a text analysis of the Union Budgets of India for financial years 2019-20 to 2023-24 highlights that the top words used in the budget speeches are “India”, “Government”, “Infrastructure” and “Sector” (Makwana, 2024).

In recent years, the intersection of digital content and predicting trends and patters has become a prevalent topic. Extensive work has been undertaken on Sentiment Analysis, highlighting its critical role in text analysis. However, it is important to acknowledge that positive sentiment does not always align with the arguer's stance, leaving room for interpretation and potential misunderstanding (Monroe, 2022).

A recent sentiment analysis study on the union budget speeches for five years from 2017-2022 highlights that the budget speeches are quite positive in sentiment (Kaushal, Ghalawat, & Saroha, 2021). Additionally, the study also shows that the key words used in the budget speeches were ‘tax’, ‘investment’, ‘infrastructure’ and ‘digital’.

Union Budget may not be utmost important in understanding the policy landscape of India and instead specific state budgets holds more relevancy (Debroy & Sinha, 2023). However, the Union budget depicts the broad fiscal agenda of the government which then forms guiding directive for the states. It is therefore crucial to analyse the speech for its complexity, readability and overall sentiment value.

3 Budget data

For the study, the official transcripts of Union Budget presented by the Indian Finance Ministers was scraped from the website hosted by National Informatics Centre and owned by Ministry of Finance, Government of India[1]. As the transcripts were in PDF format, PyPDF2 open-source Python library was used to read the text and later embedded into a data frame saved as a JSON file.

The data frame was then further processed by adding a 'year' field which mentions the year the budget was announced and 'INC_rule' field which highlights if the budget was announced during the INC rule.

The Union Budget presented each year in the Parliament from 1948 to 2024 (77 years) formed the main corpus. The below table shows the top five rows of the corpus summary.

Table 1: Top five rows of corpus summary

Text	Types	Tokens	Sentences	Sl_No	year	INC_rule
bs194748.pdf	2388	10296	358	1	1948	1
bs194849.pdf	3154	15888	618	2	1949	1
bs194950.pdf	2571	12002	509	3	1950	1
bs195051.pdf	2254	10776	402	4	1951	1
bs195152.pdf	2718	13508	472	5	1952	1

4 Methods

The various methods provided by the Quanteda-package (Benoit K. K., 2018) is used for analysis of the corpus. Firstly, the corpus was tokenised and cleaned followed by collocation and lemmatisation method to conduct statistical analysis like mean token-type ratio and frequency analysis.

Later, text complexity analysis was done followed by sentiment analysis by using an imported dictionary. Each method is further elaborated in details below.

[1] Budget Speeches | Union Budget (indiabudget.gov.in)

4.1 Stage 1: Cleaning, tokenisation and collocation

As first step, the texts field was cleaned before breaking into tokens. It is necessary to get rid of any special symbols, signs, punctuations, numbers and additional spaces. Then, the tokens were transformed to lower case for uniformity. Post that, the English language stop-words function offered by Quanteda was used to get rid of any additional preposition, articles and other filler words in English sentence. In addition to this, some other contextual words frequently found in the Union budget like India, budget, centre, etc. were removed2. Collocations were identified, assessed and then built using the textstat_collocations()

4.2 Stage 2: Corpus statistics (Lexical diversity, readability)

The mean type-token ratio test statistics was used to assess the complexity of the text. The token type refers to each unique token present in a text and therefore the ratio helps understand at what rate new types of tokens are found in a sequence of text. The mean type-token ratio for all the entire corpus was calculated and then for Union Budgets presented by INC government and non-INC government was calculated.

For assessing the readability of the text, the Flesch-Kincaid common scale was used which estimates readability in terms of school grade (years) (Flesch, 1948). Again, for readability, the score for each government type (INC and non-INC) was calculated.

4.3 Stage 3: Frequency analysis

The key words mentioned in the Union Budget were determined in terms of their relevant frequencies. A word cloud was also formed which is an easier way to understand how frequently the words are mentioned in a text. Where, the more frequent a word is, the larger the size of the text.

4.4 Stage 4: Sentiment analysis

The sentiment analysis method compared texts with a pre-defined dictionary and determined the positive and negative scores. Then, the net sentiment score is calculated by subtracting the negative from the positive score.

5 Findings

5.1 Corpus statistics

The overall lexical diversity (mean type-token ratio) of the corpus is found to be 0.22 with no major difference between the mean lexical diversity for the Union Budget speech for the INC government and non-INC government. However, if we see over time the lexical diversity of the budget speeches of the INC

[2] The complete list of super words that were removed can be found in the data analysis R script in the GitHub repository

government has increased more than the non-INC government (which is primarily the BJP-led government).

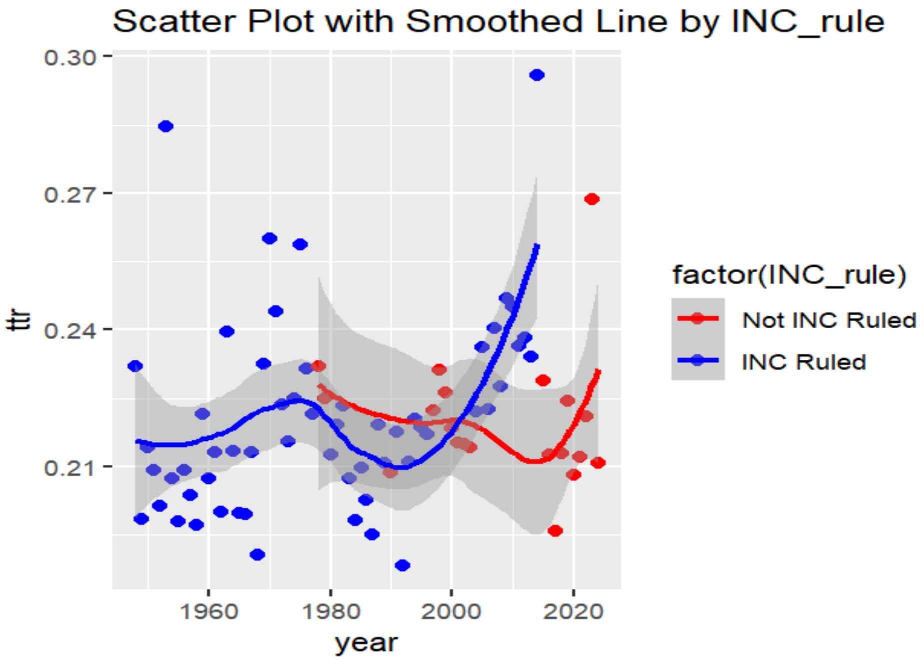

Figure 1: Scatter plot depicting lexical complexity over time

Next, coming to readability which is measured by using the common Flesch-Kincaid scale, we see that the mean readability of the Union Budget by government type doesn't vary much and is slightly more for INC government at around 12.5. This means to understand the Budget speeches the understanding and acumen level of a 13th grade student is required which in Indian context is a graduate level student.

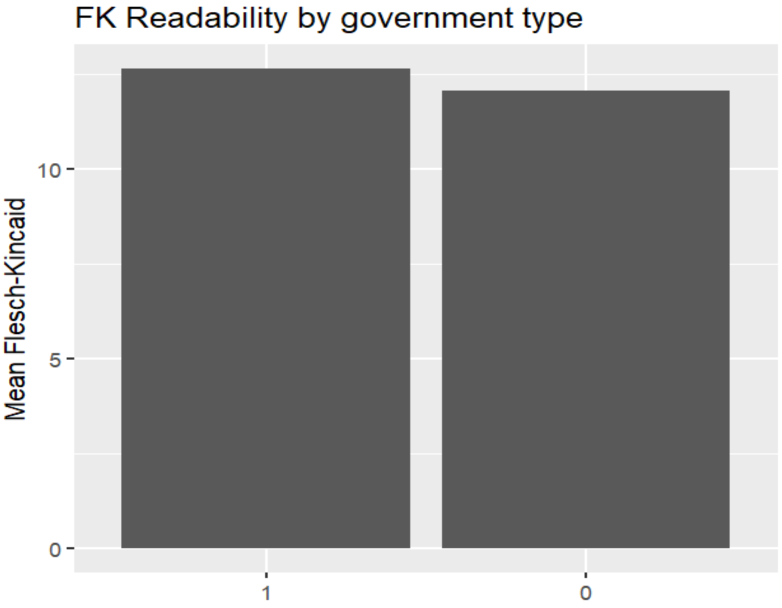

Figure 2: Readability levels by government type

5.2 Word frequency

The most pertinent is the frequency of words used in the Union Budget speeches because it reflects the main agenda of the government and the fiscal roadmap of the country for that particular year. It can be seen from the word cloud that "tax" is the most frequently used word which is not surprising because the Union Budget will most likely mention about the various tax revisions that the government intends to make for the fiscal year.

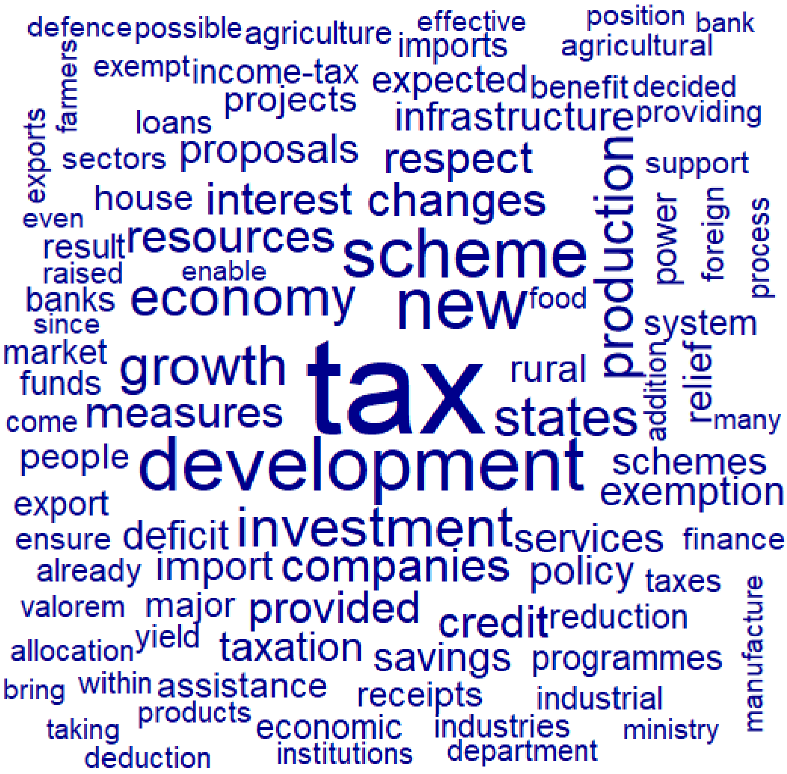

Figure 3: Word cloud of budget speeches

The second most frequently used word is 'development' which is important in the context of a developing country like India. This is followed by the word 'new' and 'scheme' which again corroborates the fact that the new economic reforms through launch of schemes for the development of the nation has always been a top priority of any government.

Other notable words are 'investment', 'economy', 'states' and 'growth' which all highlights positive economic measures that the government plans for the country.

The below graph depicts the list of most frequent words used in the Union budget speeches.

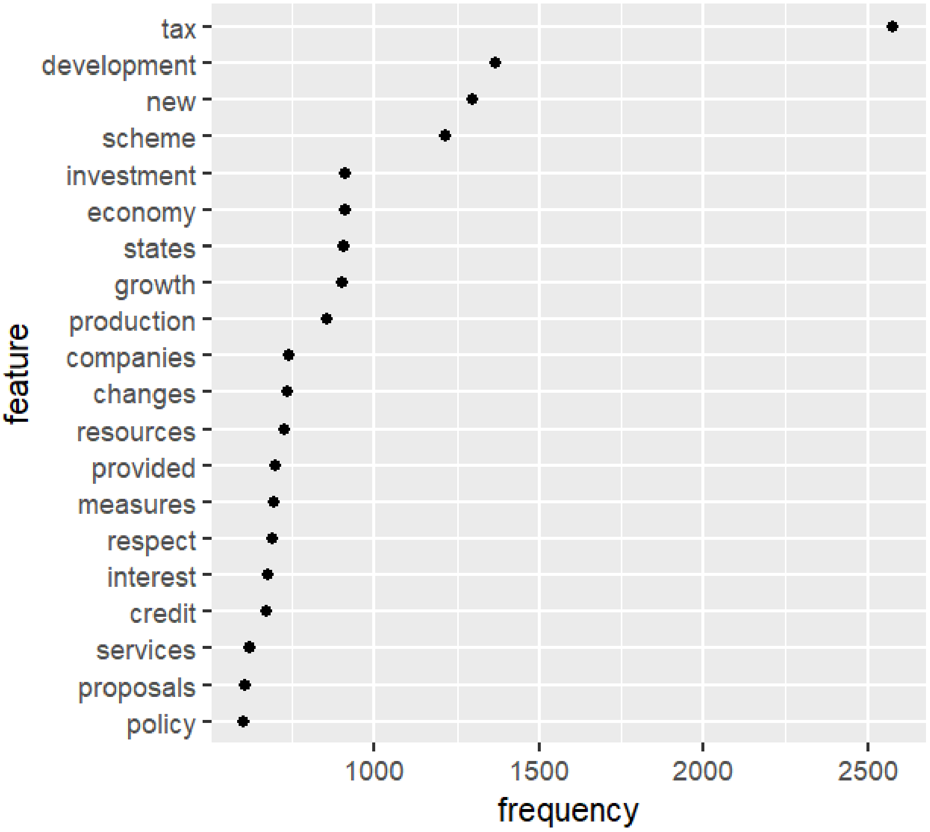

Figure 4: Frequency count of top words

5.3 Sentiment analysis

The word frequency indicated a positive sentiment of the budget speeches and it is also reflected in the net sentiment analysis graph. We can see over time how the net sentiment has increased over time. The positive tone in budget speeches since 1947 reflects a strong connection with the country's economic growth, showing a consistent and continuing optimistic outlook.

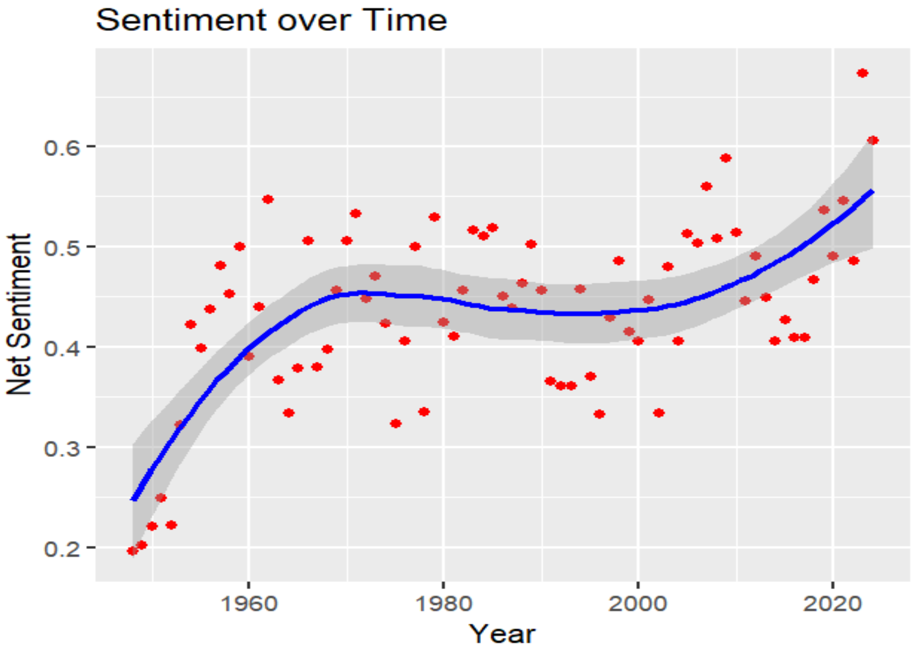

Figure 5: Net sentiment score over time

6 Discussion

The budget analysis provides enough food for thought for further research and analysis. It shows the natural evolution of the budget speeches over time and the differences between the type of government that is in power.

There is no major difference in lexical complexity or readability of the budget speeches based on government type despite the differences between their political ideologies. Moreover, the frequency of words shows that new economic reforms and development agenda has been the top priority of the government which is crucial for a young and developing nation like India.

Lastly, the sentiment analysis shows an overall positive result indicating that the budget speeches have consistently used positive words over time.

It would be crucial to further investigate how the general public perceives the budget speeches. Further research around gauging the public sentiment about the complexity, readability of the union budget will be interesting to understand. Furthermore, how the positive sentiment reflected in the budget speeches correlate with the general public sentiment, would be of interest to government and policymakers to guide their future narratives.

7 Reproducibility

The data files and analysis files including the graphs and table can be found in the GitHub repository for replication and reproducibility purposes.

8 Bibliography

- Benoit, K. K. (2018). quanteda: An R package for the quantitative analysis of textual data. *3*(30), 774. doi:https://doi.org/10.21105/joss.00774
- Benoit, K. M. (2019). Measuring and Explaining Political Sophistication through Textual Complexity. *American Journal of Political Science, 63*(2), 491–508. Retrieved from http://www.jstor.org/stable/45132491
- Bischof, D., & Senninger, R. (2017). Simple politics for the people? Complexity in campaign messages and political knowledge. *European Journal of Political Research,, 57*(2), 473–495. doi: https://doi.org/10.1111/1475-6765.12235
- Debroy, B., & Sinha, A. (2023, March 1). *Navigating the precarious balancing act: A critical analysis of the Union Budget.* Retrieved from India Foundation: https://indiafoundation.in/articles-and-commentaries/navigating-the-precarious-balancing-act-a-critical-analysis-of-the-union-budget/
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. doi:https://doi.org/10.1037/h0057532
- Kaushal, N., Ghalawat, S., & Saroha, A. (2021, December). Communicating five-year budgets for the Indian economy: Comparative text and sentiment analysis. *Journal of Content, Community & Communication, 14*. doi:10.31620/JCCC.12.21/11
- Makwana, K. (2024). A Textual Data Analysis of the Union Budget of India. *Indian Journal of Science and Technology, 17*(5), 478-486. doi:10.17485/IJST/v17i5.1174
- Monroe, S. E. (2022, April 22). Sentiment is Not Stance: Target-Aware Opinion. *Cambridge University Press, 31*(2). doi:10.1017/pan.2022.10
- *World Economic Outlook*. (2024, January). Retrieved from IMF: https://www.imf.org/en/Publications/WEO/Issues/2024/01/30/world-economic-outlook-update-january-2024