# Examining Public Sentiment and Key Themes in News Articles about the Central Vista Project: A Sentiment Analysis and Topic Modeling Study

*by Varun Sappa, Shrilakshmi S K and Utkarsh Gupta*

Mentors : Prof. Shantanu Dutta, Ashok Nimiwal

---

# Abstract

The Central Vista Project, a controversial redevelopment plan in India, has garnered widespread attention and sparked heated debates in the media and among the public. In this study, we present a comprehensive analysis of the discourse surrounding the Central Vista Project by employing topic modeling and sentiment analysis techniques.

Firstly, we use **topic modeling** to identify the main themes and topics discussed in the articles related to the **Central Vista Project**. We analyze a corpus of news articles published over a period of 4 years to identify the key issues and concerns related to the project.

Secondly, we conduct **sentiment analysi**s of the articles to determine the overall sentiment of the discourse surrounding the Central Vista Project. We apply various sentiment analysis techniques, machine learning-based approaches, to analyze the polarity and intensity of the sentiment expressed in the articles.

Finally, we explore how the sentiment and topics related to the Central Vista Project have evolved over time by analyzing the temporal patterns in the data. We investigate how the sentiment and topics have shifted in response to key events and developments related to the project, such as **court judgments**, **protests**, and also due to **COVID pandemic.**

Our findings provide insights into the complex and multifaceted discourse surrounding the Central Vista Project and shed light on the key themes, concerns, and sentiments expressed by different stakeholders. Our study contributes to the growing body of literature on the intersection of urban development, politics, and public opinion.

# Introduction

**Topic modeling** and **sentiment analysis** are two widely used techniques in natural language processing for understanding the underlying themes and sentiments present in a large corpus of text data. In this study, we apply these techniques to the discourse surrounding the Central Vista Project in India, **a highly controversial redevelopment plan** for the heart of New Delhi.

Our analysis is based on a corpus of 1157 news articles and opinion pieces published over a period of four years, from 2019 to 2022. We use topic modeling to identify the main themes and issues discussed in the discourse, and sentiment analysis to determine the overall polarity and intensity of the sentiment expressed in the discourse. By analyzing changes in sentiment and topic trends over time, we aim to gain insights into the complex and multifaceted discourse surrounding the Central Vista Project, and the concerns and sentiments expressed by different stakeholders.

Our article is organized as follows. In the next section, we provide a brief overview of the Central Vista Project and the controversies surrounding it. We then describe the methods and techniques used in our analysis, including topic modeling and sentiment analysis. In the following sections, we present the results of our analysis, including the key themes and issues identified by the topic modeling, and the changes in sentiment and topic **trends over time**.

# Central Vista Project and the controversies surrounding it

The Central Vista Project is a highly controversial urban development plan for the heart of New Delhi, India. The project aims to **redesign** and **rebuild** the central government district, including the iconic Parliament House, the Rashtrapati Bhavan (President's residence), and the India Gate war memorial. The project has faced widespread opposition from various stakeholders, including architects, urban planners, civil society groups, and opposition political parties.

Critics of the project have raised concerns about its cost, estimated to be over Rs. **20,000 crore (USD 2.7 billion)**, and the lack of transparency in the decision-making process. They have also raised questions about the environmental impact of the project, including the potential loss of green space and the impact on the heritage buildings in the area. The project has also been criticized for its alleged disregard for public opinion and the lack of consultation with local residents and civil society groups.

The controversies surrounding the Central Vista Project have been widely covered in the Indian media and have sparked heated debates in the public domain. In this report, we use topic modeling and sentiment analysis to gain insights into the discourse surrounding the project and the key concerns and sentiments expressed by different stakeholders.

# Methodology for our analysis

## Data Collection

We aimed to gather a comprehensive dataset consisting of articles from 15-20 national newspapers representing both left and right-leaning perspectives. Over the course of the project, around 1400 articles published between 2018 and 2022 were scraped. To accomplish this, a combination of tools and techniques were employed.

To initiate the data collection process, Beautiful Soup, a Python library designed for web scraping, was utilized along with the Google Console Search API. This combination allowed for systematic extraction of links associated with specific news media outlets. These links served as the basis for subsequent data retrieval.

For each collected link, Selenium, an automated web testing tool, was utilized to navigate to the corresponding article's webpage. Once on the webpage, Newspaper3k, a Python library specialized in extracting article content, date, and title, was integrated. This allowed for the efficient extraction of the required information from the visited pages.

By employing this approach, the project ensured a streamlined and efficient methodology for obtaining consistent and well-structured data from various national newspapers.

## Data Cleaning

Firstly, all advertisements present in the article content were effectively eliminated, ensuring that only relevant and meaningful information was retained. This step helped to enhance the quality and accuracy of the dataset by removing any distracting or misleading content.

Secondly, articles with outlier dates, specifically those written in 2008 and 2014, were identified and excluded from the dataset. This action was taken to maintain temporal consistency and ensure that the articles spanned the desired time frame (2018 to 2022).

By removing these outliers, the dataset remained focused on the intended period, avoiding any potential anomalies.

Lastly, articles that contained only a single mention of "Central Vista" were filtered out from the dataset. This decision was made based on the understanding that such articles lacked substantial content related to the project's analysis.
By excluding these articles, the dataset was refined to include more comprehensive and informative pieces for further examination.

## Data Preprocessing:

**Stopword Removal**: Compile a list of stopwords, including common English words, to filter out frequently occurring and insignificant words. Additional stopwords related to the Central Vista context can be added. Removing these stopwords helps focus on more meaningful content.

**Punctuation and Special Character Removal:** Eliminate punctuation marks, symbols, and special characters from the text. This step reduces noise and ensures that only relevant words and phrases remain.

**Contraction Expansion**: Expand contractions in the text to their full forms. This process ensures consistency and clarity in the language used in the dataset.

**Lemmatization:** Apply lemmatization to convert words to their base or dictionary form. This helps group together similar words and reduces word variations, enabling a more comprehensive analysis.

**Tokenization:** Utilize appropriate tokenization techniques to split the text into individual words or tokens. This can include techniques like whitespace tokenization or specialized tokenizers for better accuracy in the Central Vista context.

**Word Count:** Calculate the number of words in each article or document. This metric provides insights into the length and complexity of the content and can be useful for further analysis.

**Additional Filtering:** Filter out short words and words containing spaces, as they might not carry significant semantic meaning. Consider domain-specific filters or rules to remove irrelevant terms.

By following this structured preprocessing methodology, the text data related to the Central Vista topic will be cleaned, refined, and ready for further analysis, such as sentiment analysis, topic modeling,or extracting key insights.

## N-Grams

In addition to the aforementioned preprocessing steps, the following techniques are employed to identify meaningful bigrams and trigrams related to the Central Vista topic:

**Bigrams and Trigrams**: Bigrams are pairs of consecutive words that occur together frequently in the text, while trigrams are triplets of consecutive words. These combinations often provide valuable insights into the relationships between words and can capture specific phrases or expressions.

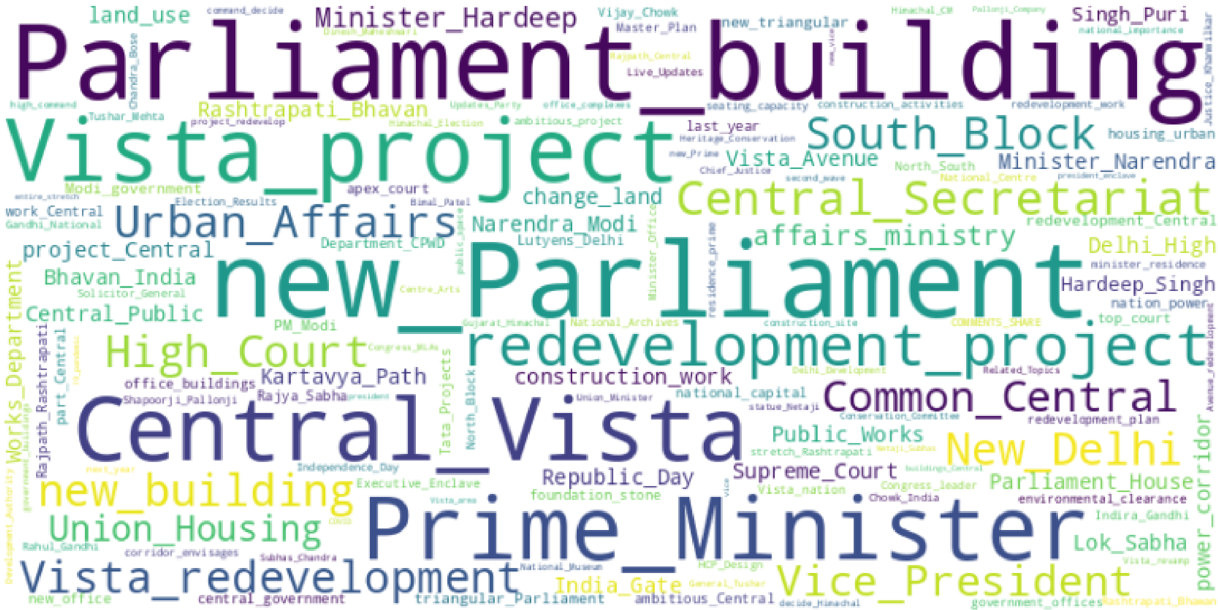
**Gensim Phraser**: The Gensim library's Phrases class is utilized to detect and generate bigrams and trigrams. The phrases object is initialized with a minimum word count of 20 for a phrase to be considered and the scoring criteria of Normalized Pointwise Mutual Information (NPMI).

**Bigram Filtering**: The bigram_counter counts the frequency of each bigram in the phrases. The top 200 bigrams with the highest frequencies are selected for further analysis. This step helps identify the most significant and frequently occurring combinations of words.

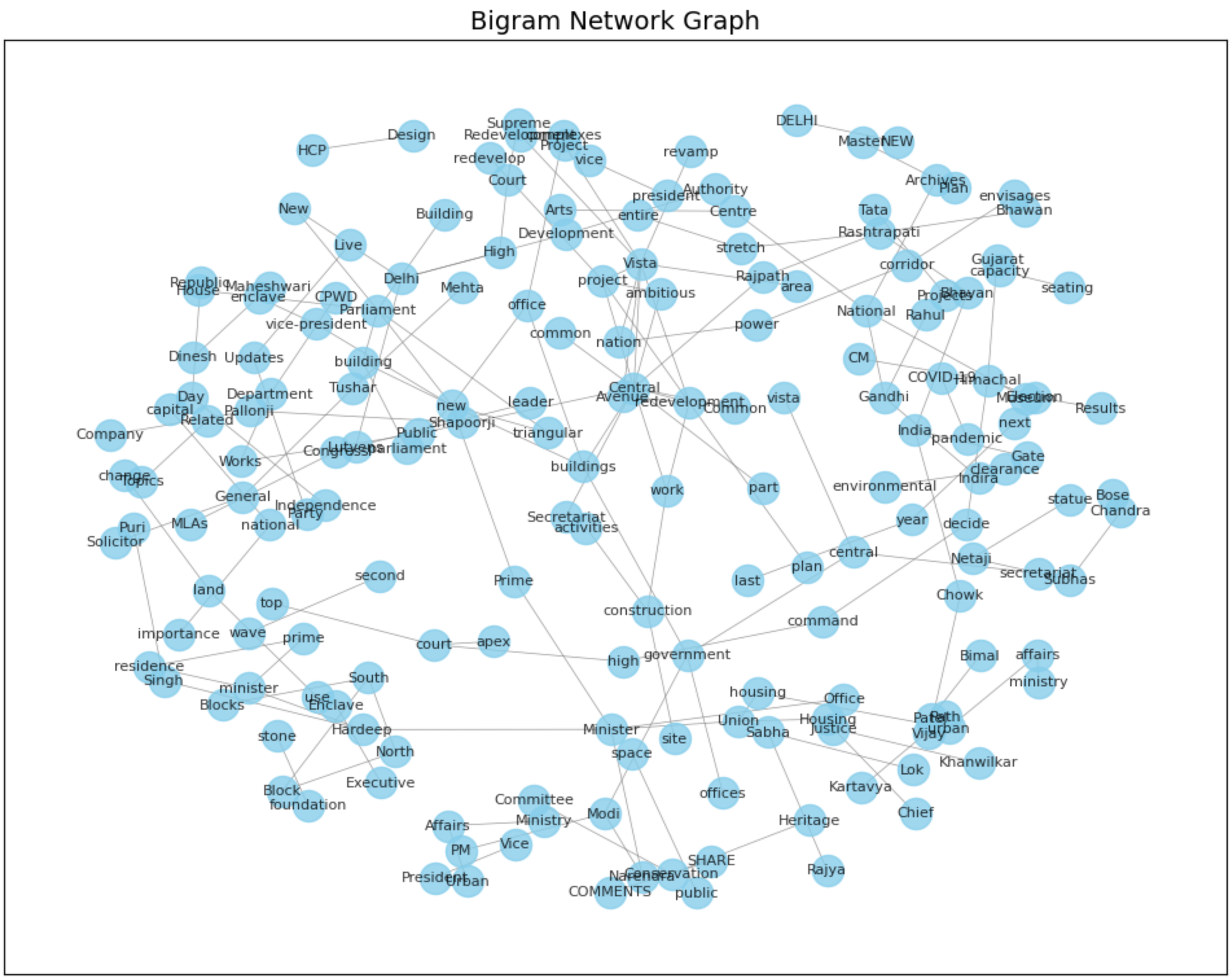
**NLTK PoS Tagging Filter:** A filter is applied to the selected bigrams using Natural Language Toolkit (NLTK) Part-of-Speech (PoS) tagging. Only the bigrams where the first word is either a noun or an adjective, and the second word is a noun are retained. This filtering is based on the observation that nouns and adjectives are more likely to represent relevant keywords in news articles. The noun_noun and adjective_noun combinations offer concise and meaningful representations of the underlying content.

**Trigram Generation**: Trigrams are obtained by combining the selected bigrams. The most common trigrams, representing frequently occurring three-word phrases, are filtered out to retain the most informative and distinctive combinations.

Bigrams




Trigrams




## NLTK collocations

The methodology for extracting relevant bigrams and trigrams related to the Central Vista topic involved the following steps:

**Bigram Extraction**: The nltk.collocations module was used to extract significant bigrams from the tokenized comments. A frequency filter retained bigrams occurring at least 50 times, and PMI scoring ranked them based on statistical relevance.

**Trigram Extraction:** Similar to bigrams, trigrams were extracted using the nltk.collocations module. A frequency filter kept trigrams occurring at least 35 times, and PMI scoring measured their statistical significance.

**Filtering Based on Parts of Speech:** Bigrams with an adjective or noun as the first word and a noun as the second word were prioritized. Trigrams with adjectives or nouns as the first and second words were considered.

**Threshold and Top-N Selection**: Bigrams and trigrams were filtered based on a PMI score threshold of 5 and selected as the top 500 associations.

**Result Analysis:** Extracted bigrams and trigrams were joined into readable formats by combining the words. Only associations with words of sufficient length were retained.

Furthermore, we performed hyperparameter tuning to optimize the results. By experimenting with different values for the parameters, we identified the most informative bigrams and trigrams.

# Topic Modeling

Topic modeling is a technique used in natural language processing to discover latent topics within a collection of documents. It involves assigning documents to different topics based on the distribution of words within them. The goal is to uncover hidden thematic patterns and gain insights into the main themes present in the dataset.

## LDA

### Data Preprocessing:

The code reads a CSV file named 'articles-corpus.csv' containing the content of all articles for topic modeling. The data is cleaned using a custom process text function, which performs the preprocessing as described in the previous section.

### N-gram Concatenation:

The code reads two CSV files named 'bigrams.csv' and 'trigrams.csv', which contain the extracted n-grams from Gensim Phraser as described in the previous section. The bigrams and trigrams occurring in the dataset are concatenated together..

### Dictionary Creation:

A Gensim Dictionary is created using the tokens from the data obtained after n-gram concatenation.

### Document-Term Matrix:

The code generates a document-term matrix (doc term matrix) using the Gensim dictionary and the tokenized text.

### Coherence Calculation and Model Selection:

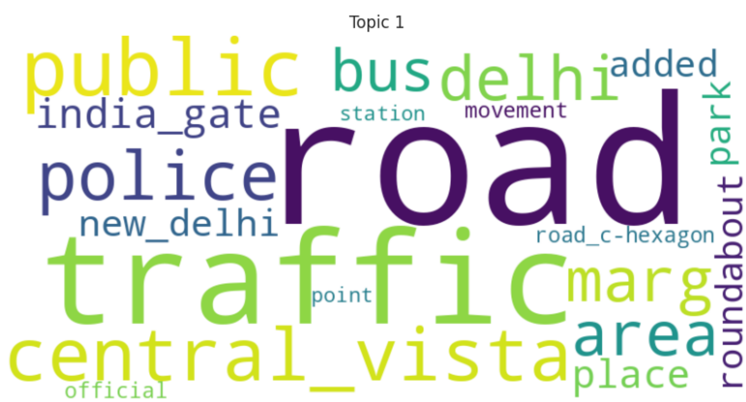
The code calculates the coherence score for different numbers of topics (ranging from 5 to 15). For each number of topics, the LDA model is trained using the doc term matrix and the Gensim Lda class. The coherence score is calculated using the CoherenceModel from Gensim. A line plot is generated to visualize the coherence scores for different numbers of topics.

### Final LDA Model Training:

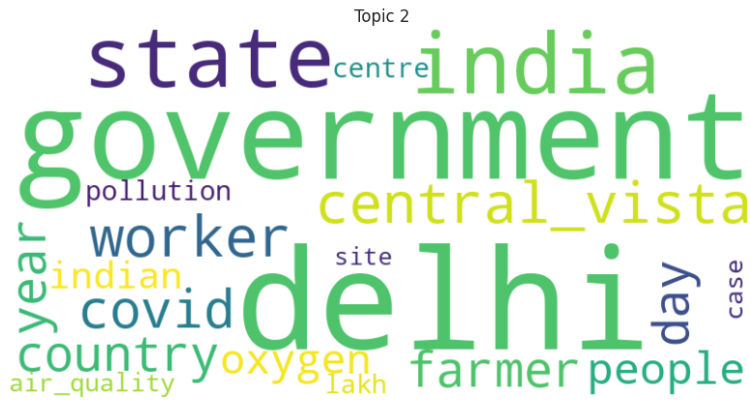
The code trains the final LDA model using 7 topics (chosen based on the coherence scores) with the Gensim Lda class.

Overall, this code performs data preprocessing, concatenates n-grams, tokenizes the text, creates a document-term matrix, selects the optimal number of topics based on coherence scores, trains the final LDA model, and displays the top words for each topic.


This topic is characterized by keywords such as "project," "petitioner," "plea," "court," and "bench." It indicates discussions related to legal aspects of the Central Vista project, including petitions, hearings, and the involvement of the Supreme Court and High Court[1] [2] [3] [4] [5]



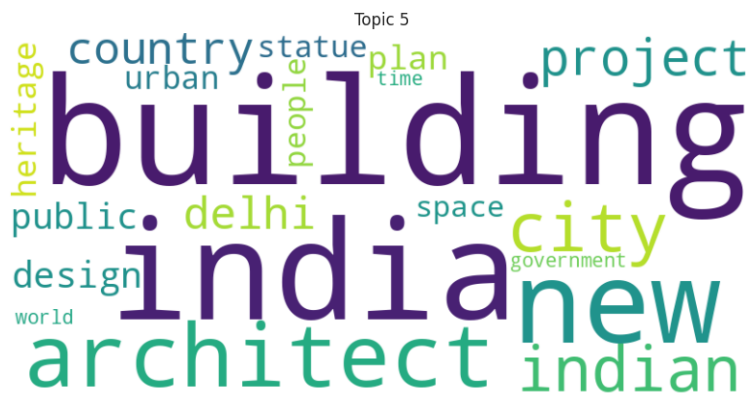

Keywords like "road," "traffic," "public," and "police" are prominent in this topic. It revolves around the impact of the project on traffic, public perception, and concerns regarding areas such as India Gate, bus routes, and design aspects of the central vista area.[1] [2 ] [3] [4] [5]



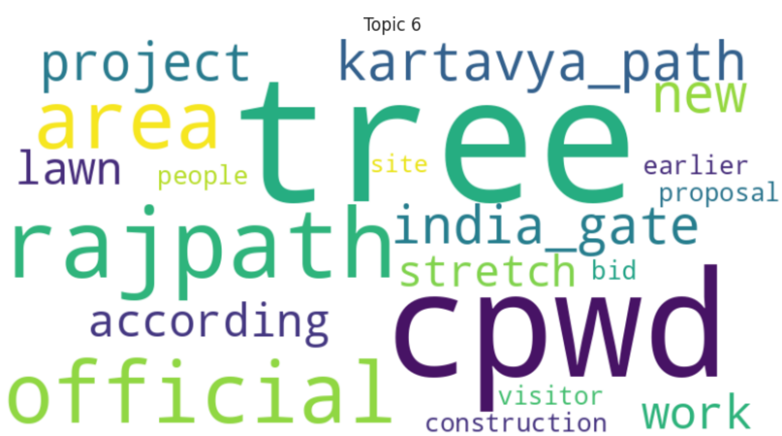

This topic focuses on broader government policies and public issues. It includes keywords like "government," "delhi," "farmer," "covid," and "worker" ,"opposition", "congress".Discussions cover various aspects such as COVID-19, workers, pollution, vaccination, farmers protest, opposition criticizing the government and the role of the government in addressing these concerns.[1] [2] [3] [4] [5]


Keywords such as "project,"“national_amblem”, “new_parliament_building” and “goverment” dominate this topic. It revolves around discussions related to the national emblem installed and debate over the faces of lions in the symbol, the construction of the new parliament building, and the involvement of different stakeholders.[1] [2] [3] [4] [5]


This topic centers around keywords like "project," "building," "government," and "plan." It discusses aspects related to the construction of various buildings, government plans, ministries, and the cost associated with the central vista project.[1] [2] [3] [4] [5]




This topic highlights keywords like "building," "india," "new," and "architect." It encompasses discussions on architectural aspects, the design of the new buildings, the cityscape of Delhi, the Indian context, and the public perception of the project.[1] [2] [3] [4] [5]



This topic focuses on keywords such as "tree," "cpwd," "rajpath," and "official." It discusses the environmental impact of the project, areas like Rajpath and India Gate, project features like Kartavya Path, construction work, and proposals related to the central vista area.[1] [2] [3] [4] [5]

These topics provide a glimpse into the key themes that emerged from the topic modeling analysis of the articles.

# NMF

Non-Negative Matrix Factorization (NMF) was used to perform topic modelling on a collection of articles and analyse the resulting topics.

Initially, the process involved the extraction of the **20 most frequently** occurring words across all articles. This was achieved by consolidating the 'Text' column into a singular list. Subsequently, the Counter class from the collections module was utilised to compute the frequency of each word, leading to the creation of a DataFrame that presents the top 20 words in conjunction with their respective frequencies.

In order to optimise the accuracy of the topic modelling procedure, a list of stop words was generated to eliminate commonly appearing words that may not substantially contribute to the topics. The stop word list was updated by incorporating words from the aforementioned DataFrame that had a frequency exceeding 1500.

Subsequently, the 'Text' column underwent **tokenization** via a tailored function that eliminated stop words (including those incorporated in the preceding stage) and restored the text to its initial form through **detokenization**. The text that underwent tokenization was subsequently saved in the DataFrame.

The determination of the optimal number of topics for the NMF model was based on the assessment of the coherence score, which is a commonly employed metric for evaluating the quality and interpretability of topics. The **NMF** model from **Gensim** was utilised, with a consideration of a range of topic numbers spanning from 5 to 20. The NMF model was trained on the tokenized texts using the **TF-IDF** representation for each topic number.

The CoherenceModel from Gensim was employed to compute the coherence score for each topic number. The aforementioned metric evaluates the semantic resemblance among words with high scores within specific topics, thereby serving as a gauge for the comprehensibility and consistency of said topics. The list **'coherence scores'** was utilised to store the coherence scores.

The optimal number of topics was determined by arranging the coherence scores of each topic in descending order and selecting the topic number with the highest score, as determined by the coherence evaluation.

A line plot was generated utilising the matplotlib library to represent the coherence scores visually. The horizontal axis was utilised to indicate the topic numbers, whereas the vertical axis was employed to denote the coherence scores. The aforementioned graph aided in determining the ideal number of topics that produced the greatest coherence score, signifying the most logically consistent and comprehensible topics for the provided dataset.

Upon determining the optimal number of topics, the Non-negative Matrix Factorization (NMF) model was applied to the tokenized texts utilising TF-IDF vectorization. The tokenized texts were transformed into TF-IDF weights using Gensim's TfidfVectorizer. The vectorizer was configured by specifying parameters such as the minimum frequency of documents, maximum frequency of documents, maximum number of features, and range of n-grams.

The NMF model was trained using the TF-IDF matrix obtained, utilising the optimal number of topics that had been determined beforehand. The Non-negative Matrix Factorization (NMF) technique was employed to factorise the Term Frequency-Inverse Document Frequency (TF-IDF) matrix into two non-negative matrices, namely the **document-topic matrix** and the **topic-word matrix.** The **'nndsvda'** initialization method was utilised along with other specified parameters to initialise the NMF model.

A topic DataFrame was generated to facilitate the interpretation and analysis of the topics. This DataFrame was composed of the most prominent words for each topic, as determined by the NMF model's outcomes.

# BERTopic

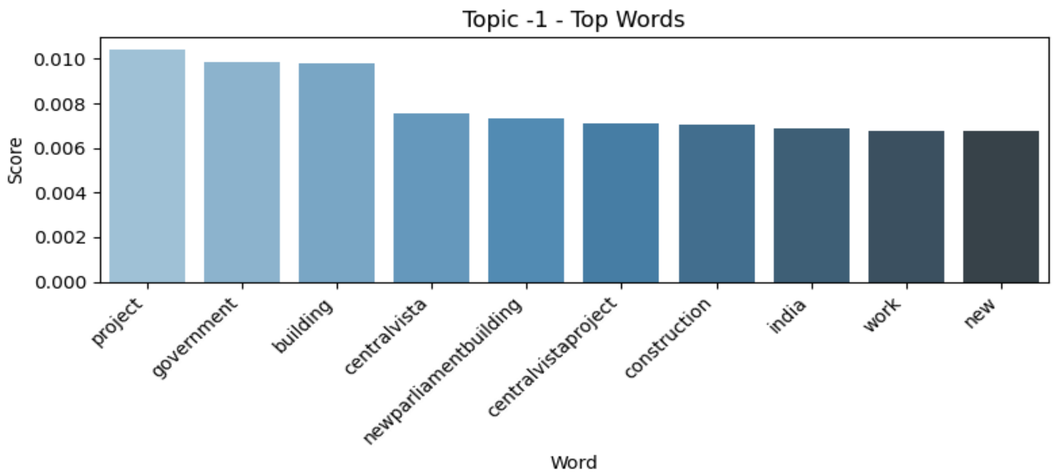
The BERTopic library was utilised to conduct topic modelling with the aim of obtaining a deeper understanding of the various topics that are present within the corpus of articles.

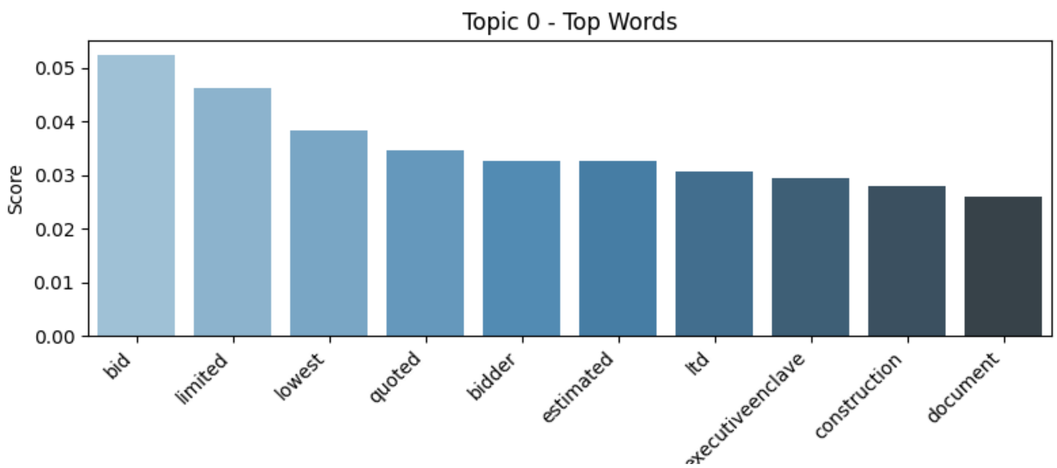
The BERTopic model was instantiated with a designated minimum topic size, which serves to establish the minimum quantity of documents necessary for a topic to be deemed legitimate. Subsequently, the model was fitted and transformed utilising the tokenized data extracted from the DataFrame.

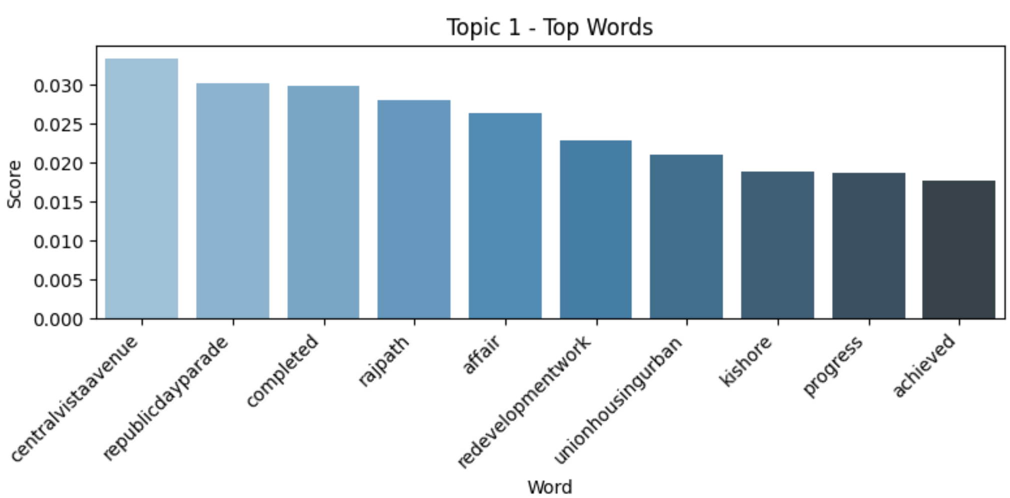
The corpus was analysed using the 'get topic freq()' method to determine the most representative topics. The output was a DataFrame that presented the topics in descending order based on their frequency of occurrence. The selection of the top 8 topics from the DataFrame was determined by the corresponding topic index and count.

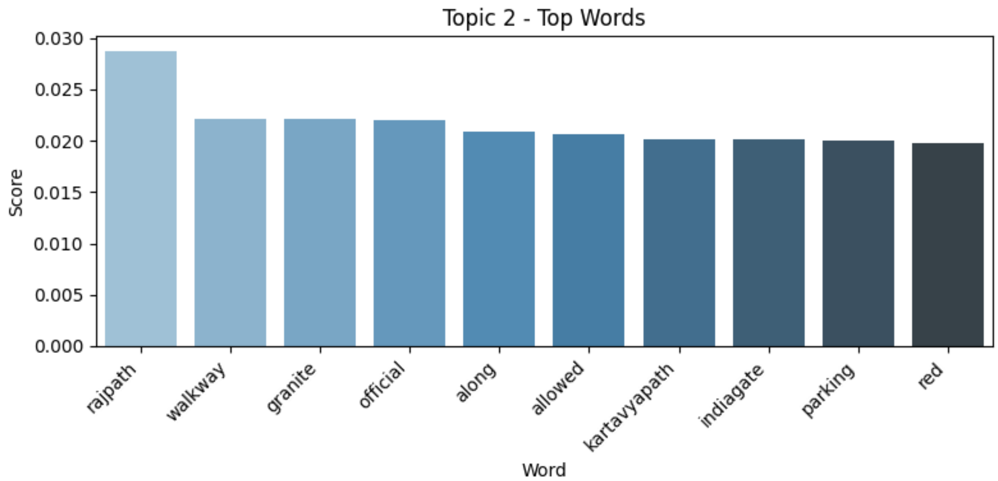
The 'get topic()' method was employed to obtain the topic words for each of the prominent topics. The method returned a list of tuples comprising a word and its corresponding score. The topic index, the number of associated articles, and the topic words were presented in print to furnish a comprehensive summary of each topic.

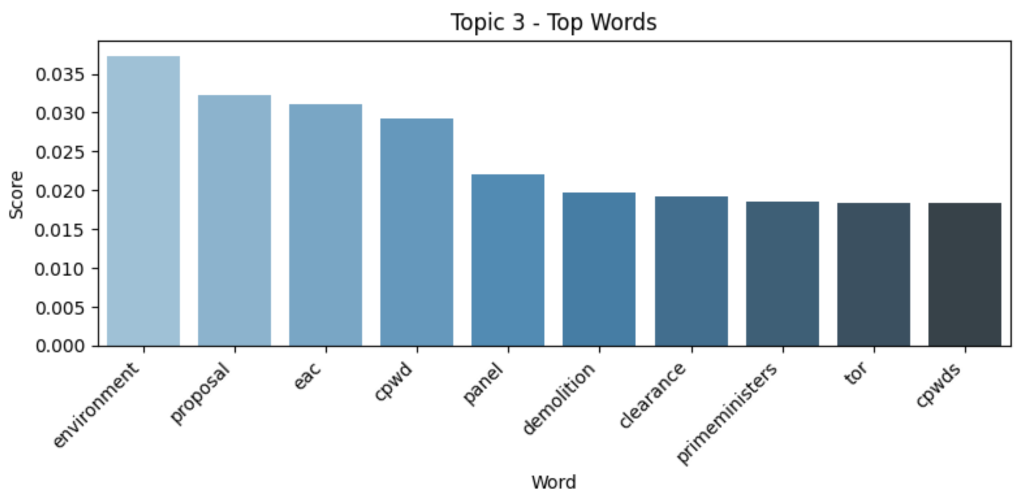
The visual representations provide a holistic perspective of the predominant terms associated with each subject matter, facilitating an enhanced comprehension of the fundamental motifs inherent in the compilation of articles. Through the examination of word distribution and scores, significant insights can be obtained regarding the primary terms linked to each subject matter, thereby facilitating subsequent analysis and interpretation.

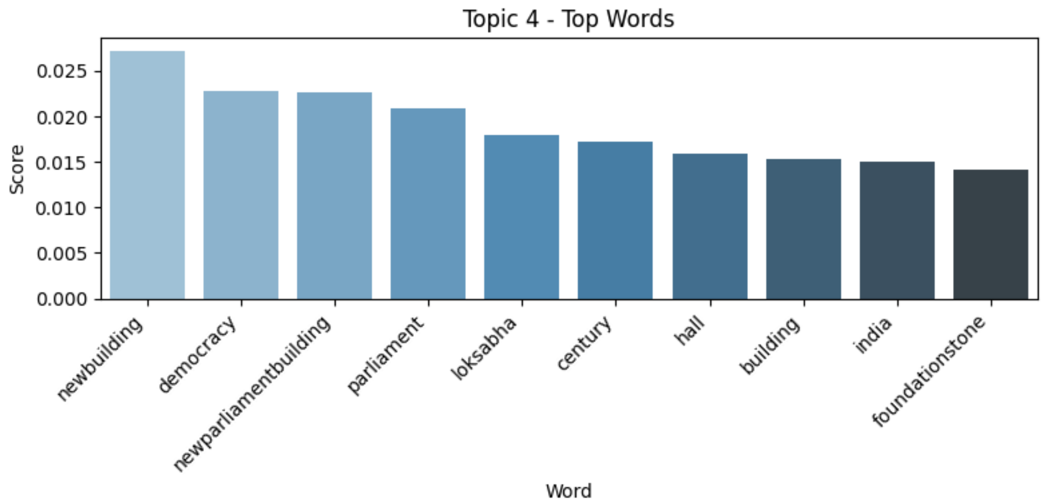

Topic -1 - Top Words
Score
0.000
0.002
0.004
0.006
0.008
0.010
project
government
building
centralvista
newparliamentbuilding
centralvistaproject
construction
india
work
new
Word


Topic 0 - Top Words
Score
0.00
0.01
0.02
0.03
0.04
0.05
bid
limited
lowest
quoted
bidder
estimated
ltd
executiveenclave
construction
document

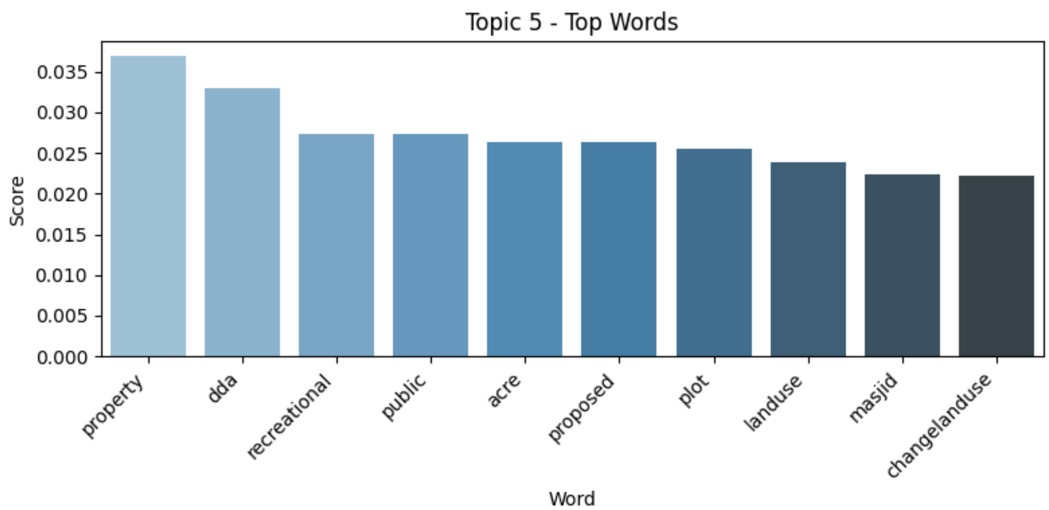

Topic 1 - Top Words
Score
0.000
0.005
0.010
0.015
0.020
0.025
0.030
centralvistaavenue
republicdayparade
completed
rajpath
affair
redevelopmentwork
unionhousingurban
kishore
progress
achieved
Word


Topic 2 - Top Words
Score
0.030
0.025
0.020
0.015
0.010
0.005
0.000
rajpath
walkway
granite
official
along
allowed
kartavyapath
indiagate
parking
red
Word


Topic 3 - Top Words
Score
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
environment
proposal
eac
cpwd
panel
demolition
clearance
primeministers
tor
cpwds
Word



Topic 4 - Top Words
Score
0.000
0.005
0.010
0.015
0.020
0.025
newbuilding
democracy
newparliamentbuilding
parliament
loksabha
century
hall
building
india
foundationstone
Word

Topic 5 - Top Words
Score
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
property
dda
recreational
public
acre
proposed
plot
landuse
masjid
changelanduse
Word

# Sentimental Analysis

Once the topics of interest were identified, sentiment analysis was performed on the articles to gain insight into the overall tone of the text. Sentiment analysis proved to be a powerful technique that uses Natural Language Processing (NLP) techniques to classify the positive, negative, or neutral sentiment of each article. This analysis provides valuable insight regarding public opinion, customer feedback, and any other text-based data where emotion played a significant role.

Using a combination of pre-trained models and custom functions, sentiment analysis was carried out. The text data was preprocessed to eliminate any irrelevant or noisy information that could have influenced the results of the sentiment analysis. Each article underwent the preprocessing function, which removed URLs, mentions, hashtags, and non-alphanumeric characters. This ensured that only relevant text was considered during the sentiment analysis.

The text was then encoded using a pre-trained tokenizer after preprocessing. Tokenization broke down the text into smaller units, such as words or subwords, enabling it to be input into the sentiment analysis model. The tokenizer utilized in this instance was based on the "cardiffnlp/twitter-roberta-base-sentiment-latest" model.

To meet the model's input requirements, the maximum number of tokens per article was capped at 512. If an article exceeded this limit, the token size was adjusted to fit within the specified length.

Specifically, the "cardiffnlp/twitter-roberta-base-sentiment-latest" model was used to conduct the sentiment analysis. This model had been trained on a large corpus of textual data and was able to accurately predict sentiment based on the provided input. The output of the model consisted of sentiment scores, which represented the likelihood of each article belonging to the positive, negative, and neutral sentiment classes.

Softmax was applied to the model's output logits to generate more interpretable sentiment scores. Softmax transformed the scores into a probability distribution, ensuring that the sum of the scores equaled 1. This facilitated a meaningful comparison of the sentiment strengths across different articles.

The sentiment scores were then added to the DataFrame, providing a comprehensive overview of the sentiment expressed in each article. Separate columns were created to store the scores for positive, negative, and neutral sentiments.

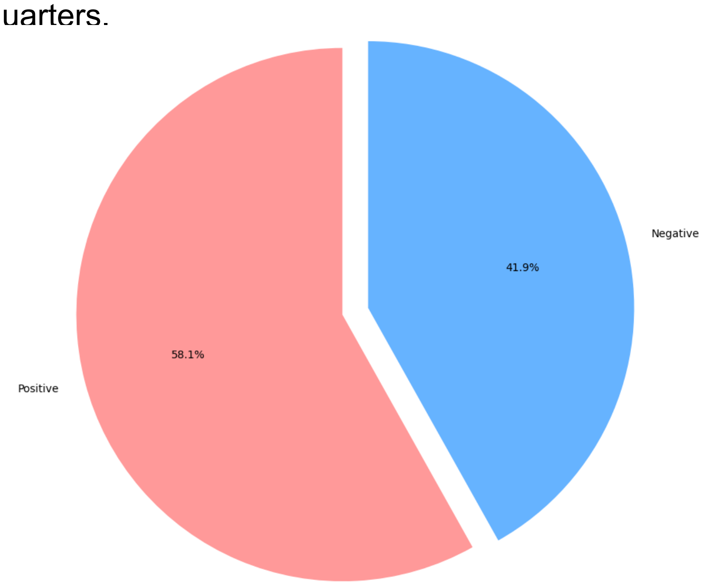
After analyzing the sentiment of the articles, the next step involved conducting time analysis to determine how the sentiment of the articles had changed over time using the publication dates of the articles. This analysis provided valuable insights into the temporal dynamics of sentiment and allowed for the identification of patterns and trends.

Monthly and quarterly sentiment scores were aggregated to perform the time analysis. Initially, sentiment scores were compiled on a monthly basis in order to identify patterns and trends in sentiment over time. Due to the short duration of a month, it was challenging to identify meaningful trends within such a brief time frame.

To overcome this limitation and capture more robust sentiment trends, the analysis was refined by grouping the sentiment scores on a quarterly basis. By extending the time frame to three months, it became easier to identify more substantial changes and trends in sentiment. Quarterly aggregation provided a better representation of the dynamics of sentiment over a longer period of time, allowing for a more accurate understanding of sentiment fluctuations.

The quarterly sentiment scores were determined by taking the median of the monthly sentiment scores. Median values were used to mitigate the influence of potential outliers and provide a more representative measure of sentiment during each quarter. This aggregation process generated a quarterly sentiment score for each of the three sentiment categories (positive, negative, and neutral).

A line plot was generated to illustrate the sentiment trends over time. The x-axis represented the quarters, while the y-axis represented the sentiment count. Separate lines were plotted for positive and negative sentiments, facilitating a direct comparison of their trends. The line plot helped identify any significant changes or shifts in sentiment across successive quarters.



# Sentimental Analysis Over Time

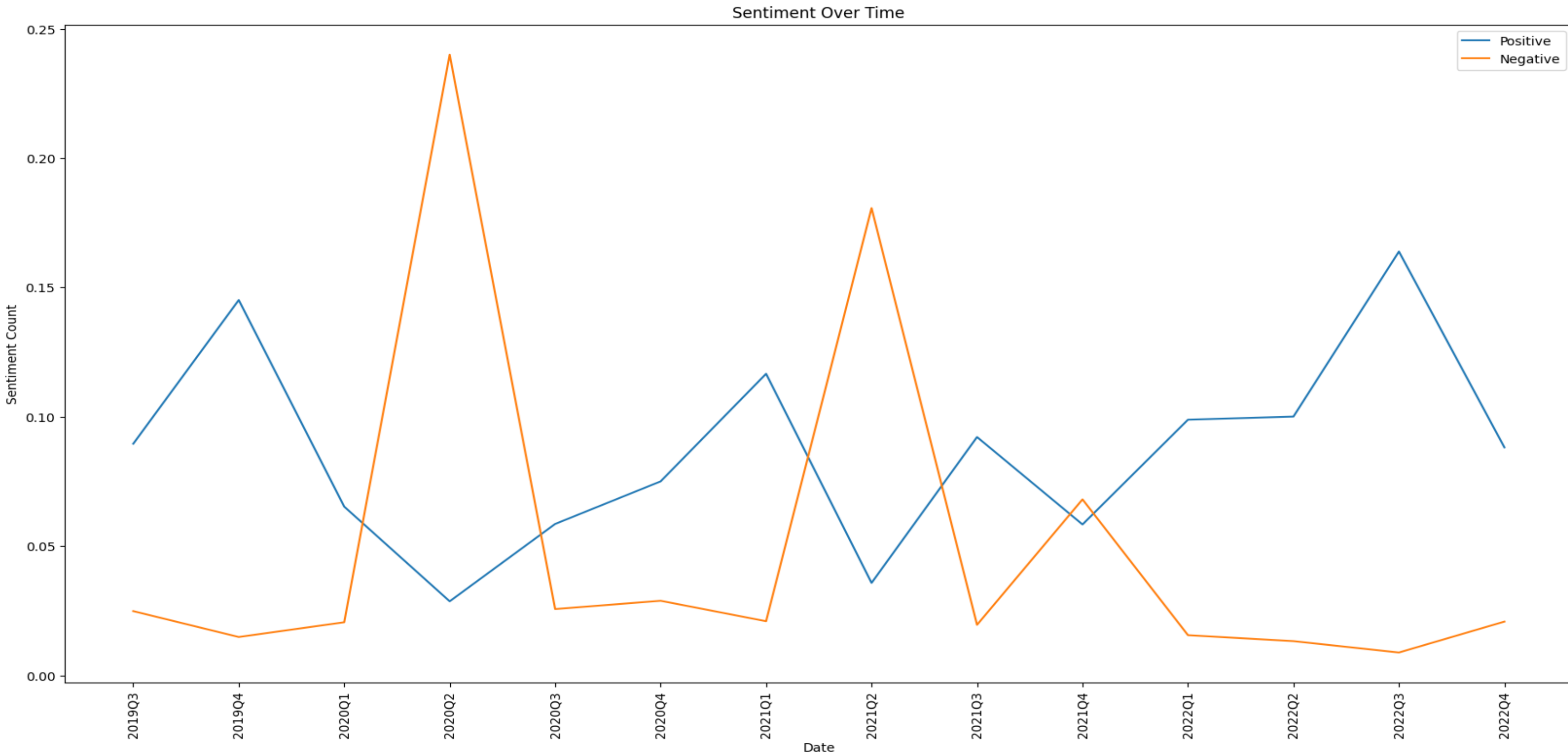
The objective was to analyse sentiment trends over time for particular topics obtained through topic modelling. The strategy consisted of a series of data preprocessing steps, aggregation techniques, and visualisation methods for deriving insightful information from the data.

In the first step of the methodology, the dataset was filtered using a custom filtering function based on the assigned topic probability. The topic was assigned to articles whose scores surpassed a predetermined threshold. This filtering procedure assisted in isolating articles that were specifically relevant to the topic of interest, and the resulting subset of data was saved in a DataFrame titled 'topic'.

Next, quarterly summaries of the sentiment scores for each article in the 'topic' DataFrame were generated. The quarterly aggregation involved grouping the data according to the 'quarter' column, which was derived from the 'joined' column, which represented the publication date of the articles. The sentiment scores for each quarter were then replaced with the median sentiment scores, which provided a more representative measure of sentiment for each period.

Using the matplotlib library, a line plot depicting the sentiment trends over time was produced. This line graph displayed the sentiment trends over time, providing insight into the changes in sentiment for the selected topic over the specified time period.

This research methodology provided a methodical method for analysing sentiment trends over time for a particular topic. By employing data filtering, sentiment analysis, and quarterly aggregation techniques, it is possible to gain valuable insights into the evolution of sentiment. The line plot visualisation enhanced the results' interpretability, allowing us to identify patterns, trends, and sentiment shifts over time for the selected topic.




The general sentiments of news articles on the Central Vista Project have displayed an interesting trend over time. As the project has progressed, there has been a noticeable increase in positive sentiments surrounding it. This can be attributed to several factors, including the implementation of development plans, infrastructure improvements, and the positive economic impact expected from the project.

However, it's important to note that there have been exceptions to this overall positive trend. For instance, during the peak of the COVID-19 pandemic, negative sentiments spiked due to concerns about public health and the allocation of resources towards the project instead of healthcare infrastructure.

Furthermore, the project has also faced opposition and protests from certain groups, resulting in temporary spikes of negative sentiments. These instances of protest and public dissent have led to increased criticism and negative portrayal of the project in some news articles.

Nonetheless, despite these occasional fluctuations, the general trajectory has shown a decrease in negative sentiments over time. This can be attributed to a combination of improved public perception, the realization of project benefits, and the positive impact on urban development.

# Detailed Timeline analysis of Sentiment Analysis Graph

**2019 Q3**

- GoI considers all options available to resolve the New Parliament debate [1]
- GoI finalizes the decision of construction of New Parliament and conceives the master plan of Redevelopment of Central Vista Avenue [1] [2]
- Top reputed architect firms interested in Central Vista Redevelopment Project [1]
- GoI promises to not destroy heritage buildings [1]
- Opposition questions the necessity of new parliament [1] [2]
- Environmentalists are concerned about the implications of the project on biodiversity [1]

**2019 Q4**

- Architect firms submit bids [1]
- GoI promises no deforestation or destruction of heritage buildings for the Central Vista project [1] [2]
- Ahmedabad-based Firm Wins Consultancy Bid for Central Vista Revamp Amid Accusations of Modi Govt's Partiality [1]
- Revamped Rajpath to feature Eiffel Tower-like structure [1]
- Workers express pride as PM Modi acknowledges contribution to Central Vista project [1]

**2020 Q2**

- 70 former bureaucrats call for halt to Central Vista revamp amid COVID-19 pandemic [1]
- Opposition calls to scrap Central Vista project to fund COVID-19 relief [1] [2]
- The Supreme Court receives petitions against the Central Vista project [1]
- GoI receives backlash over DA (Dearness Allowance) freeze [1]
- Central Vista revamp lacks public consultation, expert advice, environmental norms [1]

**2020 Q3**

- GoI increases transparency of Central Vista plans [1]
- BJP leaders address the necessity of Central Vista [1] [2] [3]
- Seven companies bid for the new Parliament building, signaling strong interest [1]

**2020 Q4**

- SP Group withdraws letters alleging irregularities in new Parliament building bid [1]
- CPWD amends application for environmental clearance, adding environmental safeguards [1] [2]
- Central Vista project will save money and improve efficiency, government tells SC [1]
- The foundation stone-laying ceremony for the Central Vista project takes place [1] [2]

**2021 Q1**

- Heritage Conservation Committee approves construction of new parliament building [1] [2]
- Supreme Court approves Central Vista project [1]
- Bhoomi Pujan ceremony is performed [1] [2]

**2021 Q2**

- First look of Rajpath buildings [1]
- Congress, Shivsena, CPI and more opposition parties accuse BJP of continuing with the expensive Central Vista project during COVID-19 pandemic [1] [2] [3] [4] [5] [6]
- Environmentalists say the Central Vista project is being broken up into smaller parts to hide its true environmental impact [1]
- PIL filed against the construction of Central Vista [1] [2] [3]
- Opposition calls for more focus on Farm Laws over Central Vista [1]

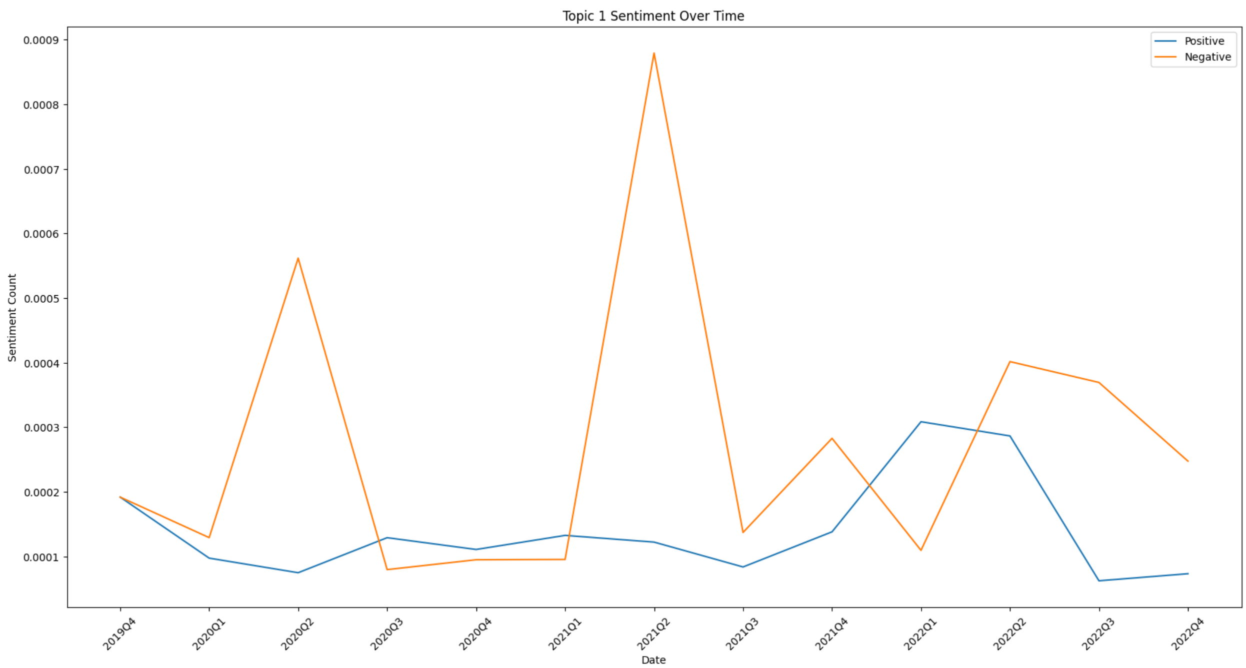
**2021 Q3**

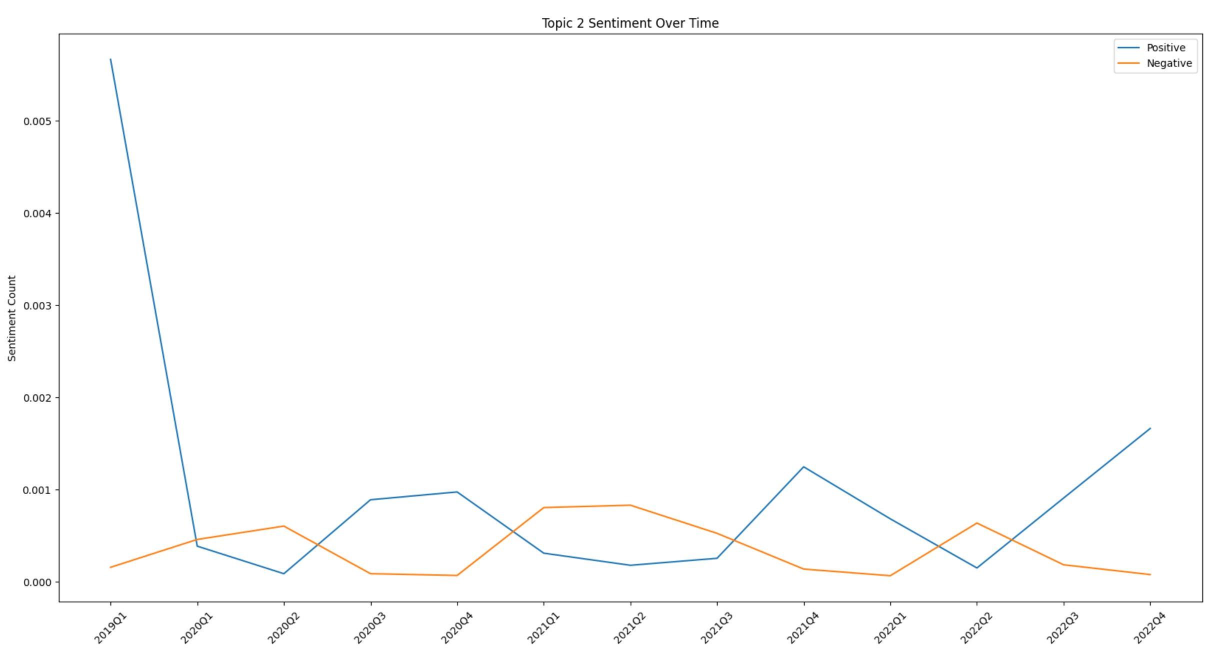
- Only Rs. 238 crore spent on Central Vista [1] [2]
- The next Republic Day parade to be held in a revamped Rajpath with many celebrities in attendance [1] [2]
- Central Vista website launched [1] [2]
- Modi inaugurates New Defence Ministry Office Complexes [1] [2] [3] [4] [5]

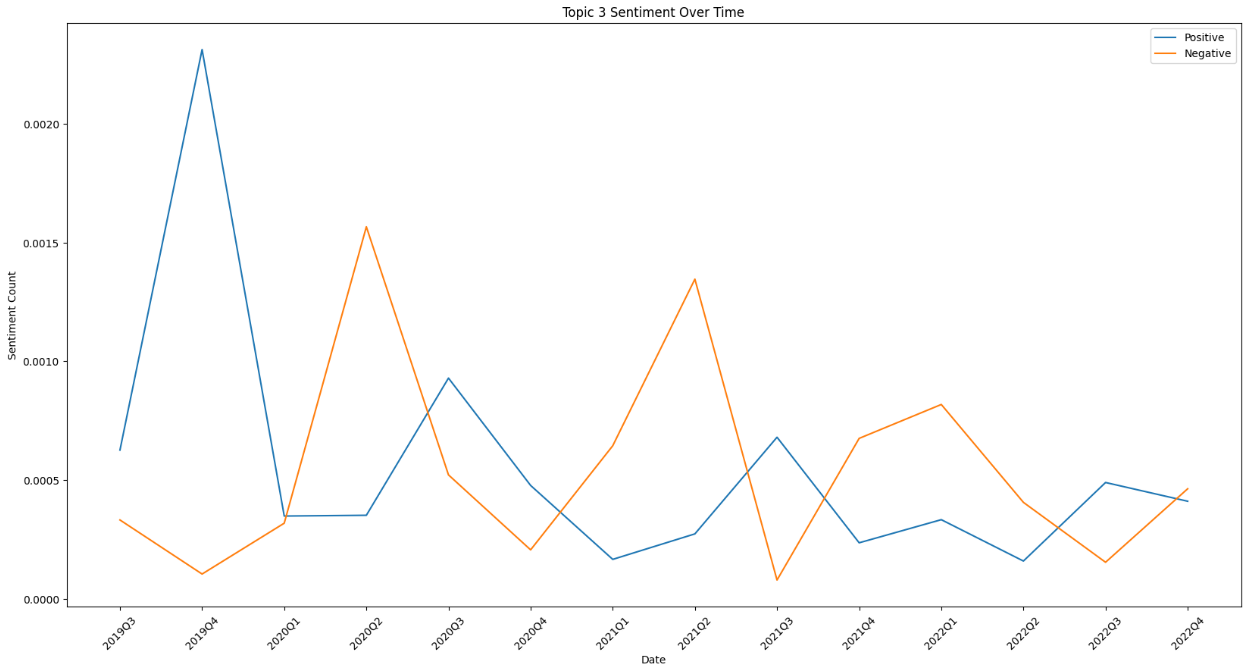
**2021 Q4**

- Biodiversity park to be constructed on Yamuna banks [1]
- GoI takes several initiatives to boost agricultural production [1]
- Republic Day will be hosted in New Parliament building [1] [2]
- Supreme Court asks the government to explain the need for land use change in the Central Vista project [1]
- GoI receives backlash over farmers' issues and suspension of MPs [1]
- Senior advocate Vikas Singh seeks to stop the Central Vista project due to dust pollution. Delhi schools and colleges shut down due to pollution [1] [2]

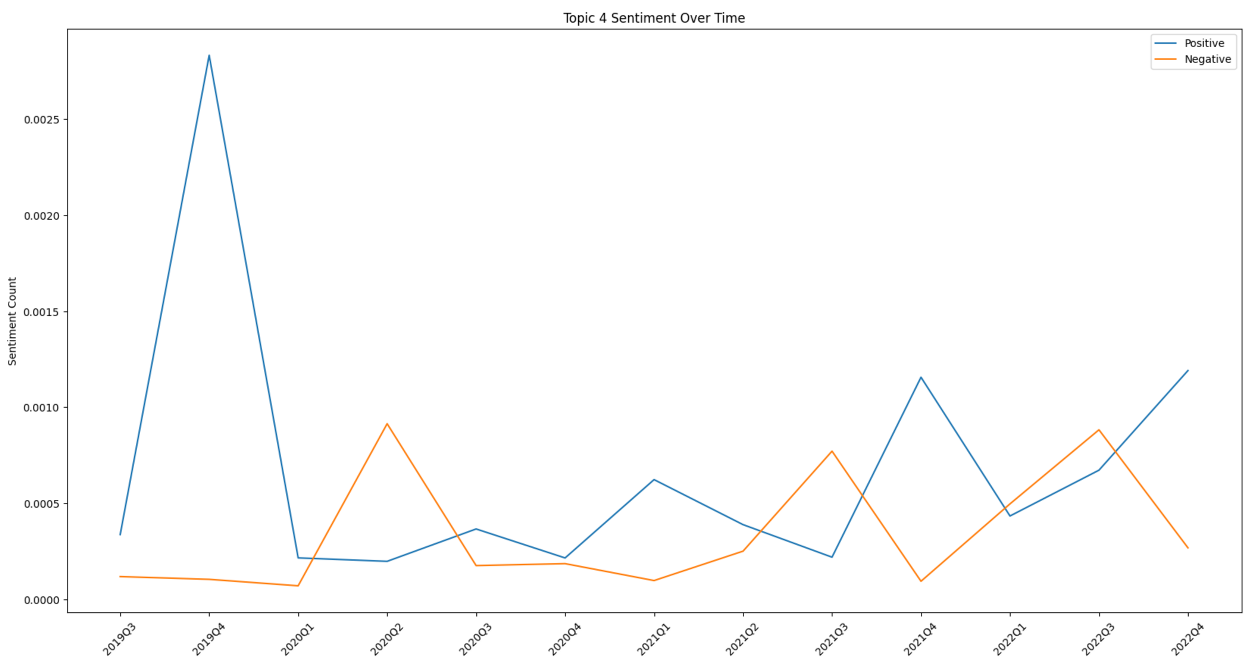
# Topicwise Sentiment Analysis



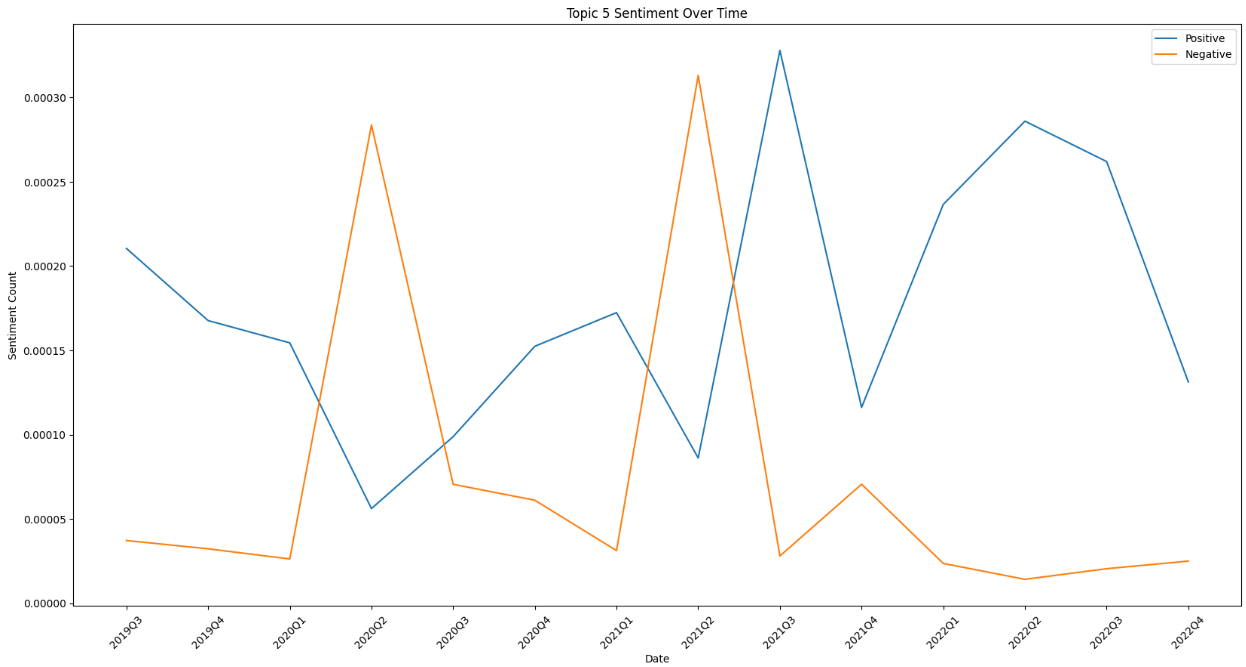



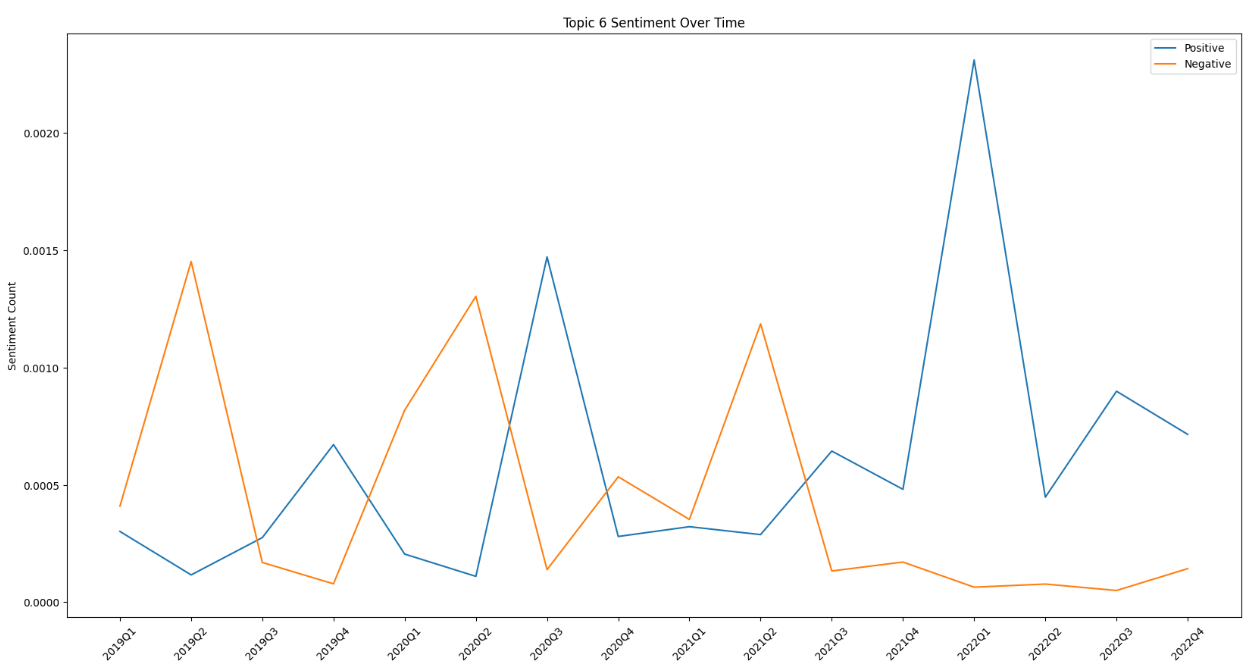


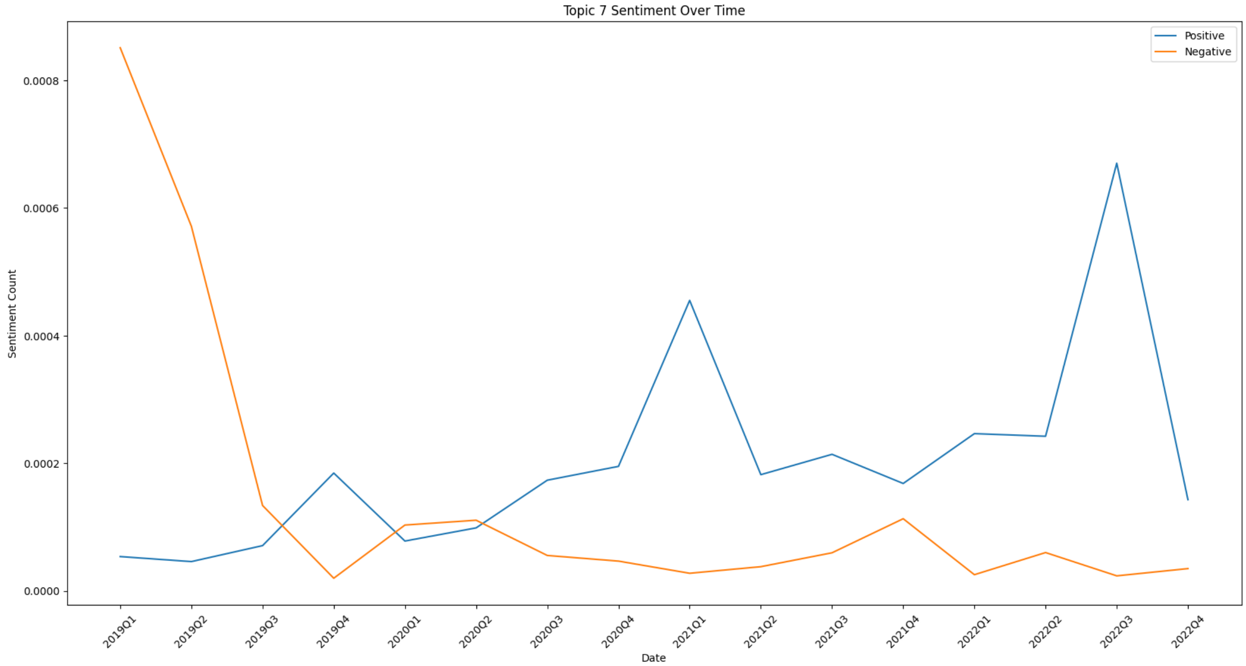
Topic 3 Sentiment Over Time
Positive
Negative
Sentiment Count
0.0000
0.0005
0.0010
0.0015
0.0020
2019Q3
2019Q4
2020Q1
2020Q2
2020Q3
2020Q4
2021Q1
2021Q2
2021Q3
2021Q4
2022Q1
2022Q2
2022Q3
2022Q4
Date

Topic 4 Sentiment Over Time
Positive
Negative
Sentiment Count
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
2019Q3
2019Q4
2020Q1
2020Q2
2020Q3
2020Q4
2021Q1
2021Q2
2021Q3
2021Q4
2022Q1
2022Q3
2022Q4
Date


Topic 5 Sentiment Over Time
Positive
Negative
Sentiment Count
Date
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
2019Q3
2019Q4
2020Q1
2020Q2
2020Q3
2020Q4
2021Q1
2021Q2
2021Q3
2021Q4
2022Q1
2022Q2
2022Q3
2022Q4


Topic 6 Sentiment Over Time
Positive
Negative
Sentiment Count
Date
0.0000
0.0005
0.0010
0.0015
0.0020
2019Q1
2019Q2
2019Q3
2019Q4
2020Q1
2020Q2
2020Q3
2020Q4
2021Q1
2021Q2
2021Q3
2021Q4
2022Q1
2022Q2
2022Q3
2022Q4




Sentiment analysis of the topics generated from topic modeling over a four-year period provides valuable insights into public sentiment towards the Central Vista Project. This analysis helps track sentiment trends, understand public perceptions, and identify key moments or events that influence sentiment. The findings assist stakeholders in gauging project reception, refining communication strategies, addressing concerns, and aligning policy decisions with public sentiment.

Additionally, sentiment analysis can be a valuable tool for understanding public sentiment towards policy decisions related to the project and guide future decision-making processes. It can highlight areas where public sentiment aligns with project objectives and identify potential areas of conflict or controversy that require further attention.

# Conclusion

In conclusion, our research and analysis utilizing advanced Natural Language Processing (NLP) techniques have yielded significant findings regarding the Central Vista Project in India. By employing topic modeling and sentiment analysis, we have successfully identified and examined seven prominent themes that have emerged from the discourse surrounding the project. Additionally, our comprehensive sentiment analysis has provided valuable insights into the diverse range of sentiments expressed by various stakeholders.

These findings hold substantial implications for policymakers, government agencies, and other relevant stakeholders involved in the Central Vista Project. The identification of key themes allows for a deeper understanding of the concerns, priorities, and perspectives of different groups. This knowledge can facilitate more informed decision-making, fostering a more inclusive and sustainable development process.

While our study has shed light on important aspects of the Central Vista Project discourse, there are avenues for future research to expand upon our findings. Further investigations could incorporate a wider array of data sources, such as social media platforms, to capture a more comprehensive representation of public opinion. Additionally, analyzing the long-term impacts of the project on various dimensions, such as social, economic, and environmental aspects, would provide deeper insights into its consequences.

In summary, our research contributes to the ongoing dialogue surrounding the Central Vista Project by providing evidence-based insights and a nuanced understanding of the multifaceted perspectives associated with it. These findings aim to inform decision-makers and stakeholders, facilitating more holistic and inclusive approaches to development and ensuring the long-term sustainability of the project.