Lok Sabha Debates Analysis

MEET SABLE, DAIICT, India
KHUSHKUMAR KANTARIA, DAIICT, India
KRISHNA PARMAR, DAIICT, India

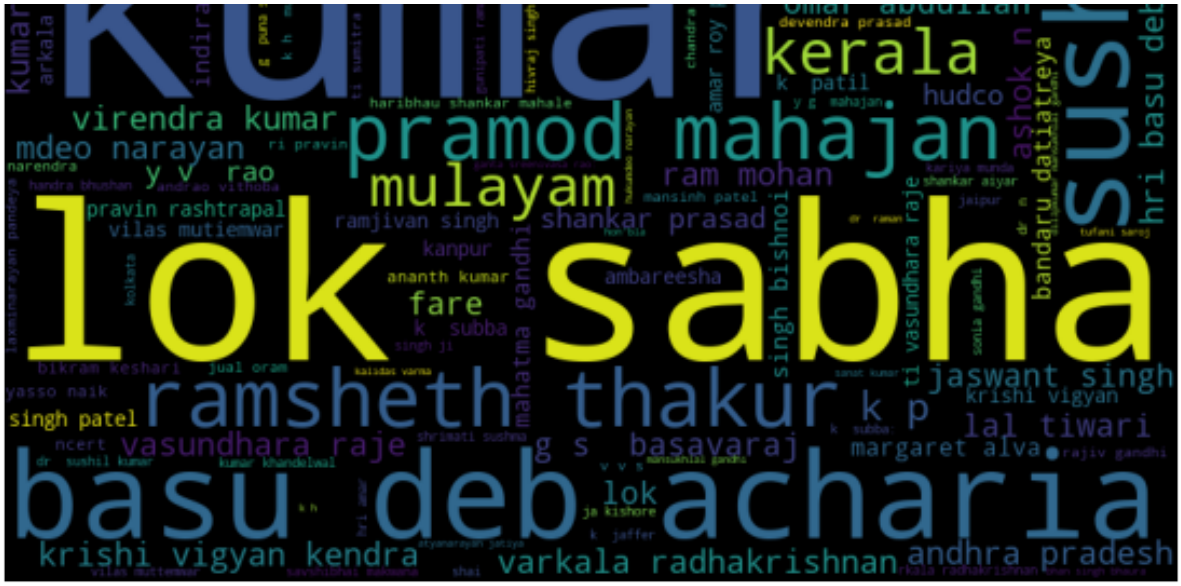


Fig. 1. Word Cloud of Lok-Sabha Debate from 2001

Legislative Assembly data set was used from the Parliament Digital Library (PDL) to analyze and differently visualize the entities and their occurrences/frequencies with the help of Named-Entity Recognition. This also helps discover the trends of topics, places, names, dates, and many more. Not only that, but the summarizer also provides a summary of all the long files of any particular year.

Keywords: Lok Sabha, Named-Entity Recognition (NER), Word-cloud, Summarizer

ACM Reference Format:
Meet Sable, Khushkumar Kantaria, and Krishna Parmar. 2022. Lok Sabha Debates Analysis. In *NLP Project: Lok Sabha Debates Analysis.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Lok Sabha Debates Analysis, December 3rd, 2022, NLP Project, DAIICT

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Natural Language Processing, commonly known as NLP, is simply the mixture of computer science and linguistics. Alexa, Siri, Google Translator, and Auto-complete, are a few NLP models that we commonly use in our day-to-day life. The main essential goal is to train the computers to learn, understand and extract the meaning of languages, the way humans do. However, in case of extensive data it becomes difficult for humans to extract the meaning out of it and get some meaningful insights. We can analyze the given data and beautifully visualize it using numerous Natural Lanaguage Processing Methodologies.

2 BACKGROUND INFORMATION

India, popular because of its diverse culture and rich politics, has a parliament as its back, which divides its power into two major houses, the Lok Sabha(the House of People) and the Rajya Sabha(Council of States). Throughout the years after Independence, many debates were held in both houses where many party members participated and projected their views, opinions, and thoughts towards the nation's situation, economics, issues, wealth, and health, directly addressing each other by name. All debates were recorded by the Secretary of India and were made publicly available to the common public residing in India.

Parliament of India, Lok Sabha, is the house of the people, chosen through direct elections by the people. As per the Indian Constitution, the maximum strength is 552, which comprises 530 representatives of different states, 20 representatives of the Union Territories, and not more than 2 from the Anglo-Indian community, nominated by the President of India if the participation is not adequate.

Narrowing down this wide perspective to the House of People, which plays an important role in politics where members talk about more state and financial issues, which is crucial in building a government.

3 METHODOLOGY

3.1 Dataset Preparation

To prepare our dataset, first we went through the Parliament Digital Library and decided a particular dataset to analyse. We decided to go forward with "Lok Sabha Debates."

3.1.1 Web Scrapper. **:** To download the files corresponding to particular year, we made a web-scraper. In process, we used beautiful-soup library from python and got all *a-tag href* links and downloaded the data. Given a input of a year, it scrapes all the pdf files of speeches of that year's Lok Sabha Debates and saves them in google drive folder for further use.

3.1.2 PDF2CSV Converter. **:** As this data is in PDF files, it cannot be processed directly. Hence, we converted the PDF files to CSV files. In process, we used tika library from python to read pdf files and used pandas to convert it to .csv. As it also contained lots of whitespace, we used re library from python to remove the whitespace. This converter takes pdf as input and gives a CSV file as output which is being saved to the google drive folder. Now, this csv can be further processed for analysis purposes.

3.2 WORDCLOUD

Word cloud, also known as Tag cloud, wordle, or weighted list in visual design, is an image showing different words in different orientations and sizes depending upon the word's frequency or importance. There are three main types of word cloud applications in social software wherein size of the word in the word cloud represents

frequency, significance, and categorization. Here, **wordcloud** python library is used for better visualization of instances with higher significance or impact.

3.3 Named-Entity Recognition (NER) Tagging

NER is a technique used for identifying or detecting the key entities from the text and classifying them into multiple pre-defined categories. NER was performed using spaCy, a free and open-source python library generally used to build information extraction and natural language models. Below table shows the description of the different categories:

Categories and their description	
Category	Description
PERSON	People, including fictional
CARDINAL	Numerics that do not fall under any other type
ORG	Companies, agencies, institutions, etc
GPE	Countries, cities, states, etc
NORP	Nationalities or religious or political groups
ORDINAL	"first", "second", "third", etc
DATE	Absolute or relative dates or periods
PRODUCT	Objects, vehicles, food, etc
TIME	Times smaller than a day
QUANTITY	Measurements as of weights or distance
LANGUAGE	Any named language
MONEY	Monetary values including units
PERCENT	Percentage, including "%"
LOC	Non-GPE locations, mountain ranges, bodies of water
EVENT	Named hurricanes, battles, wars, sports events, etc.
LAW	Named documents made into laws
FAC	Buildings, airports, highways, bridge, etc
WORK_OF_ART	Title of books, songs, etc

3.4 Summarization

Summarization is the method of computationally reducing a set of data to produce a subset that highlights the major points or information inside the source text.
Our attempt here was to summarize the data of speeches throughout the year and save it in a file. To do that, we used python libraries such as spacy, pandas, nltk. The pipeline of summarizer is as follows: As you can see in fig. 2, first we combined data from all the csv files of particular year. This data was cleaned and then tokenized. As per definition of summary above, we need some part of data which can show major points. This can be found out using sentence scores with help of word frequency table. Thus, we tokenized word and sentence of data and then further produced sentence score. Using heap, we selected 0.01% of data from each file of year as summary. This summary on average produced 10 page file.
Informed note: Due to all files being scanned in different format by government, there were some unworkable redundancy in data which gave some fluctuation in summary output which was less in amount and not much significant. Another was, as this files include speaker names per say, it also came in the summary as fluctuation which is not much significant. Quality of summary is good and readable.

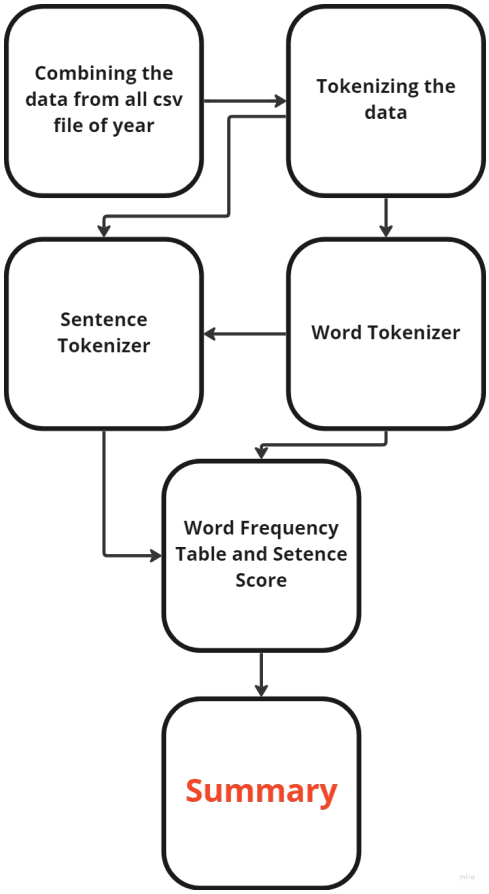


Fig. 2. Structure of summarization

3.5 DEPENDENCY PARSER

The technique of evaluating the relationships between the phrases in a sentence to ascertain its grammatical structure is known as dependency parsing. Based mainly on this, a sentence is broken into numerous components. The method is predicated on the notion that each linguistic component of a phrase has a direct link with the others. Dependencies are the names given to these relationships.

To do that, we used spaCy library and got dependency graph of a file per year. Due to file data being too big, the graph is immensely big.

4 CASE STUDY: LOK SABHA SPEECHES OF 2001

Note: Analysis can be done for any year as per user input but here we are only showing for case of 2001.

4.1 VISUAL REPRESENTATION OF NER

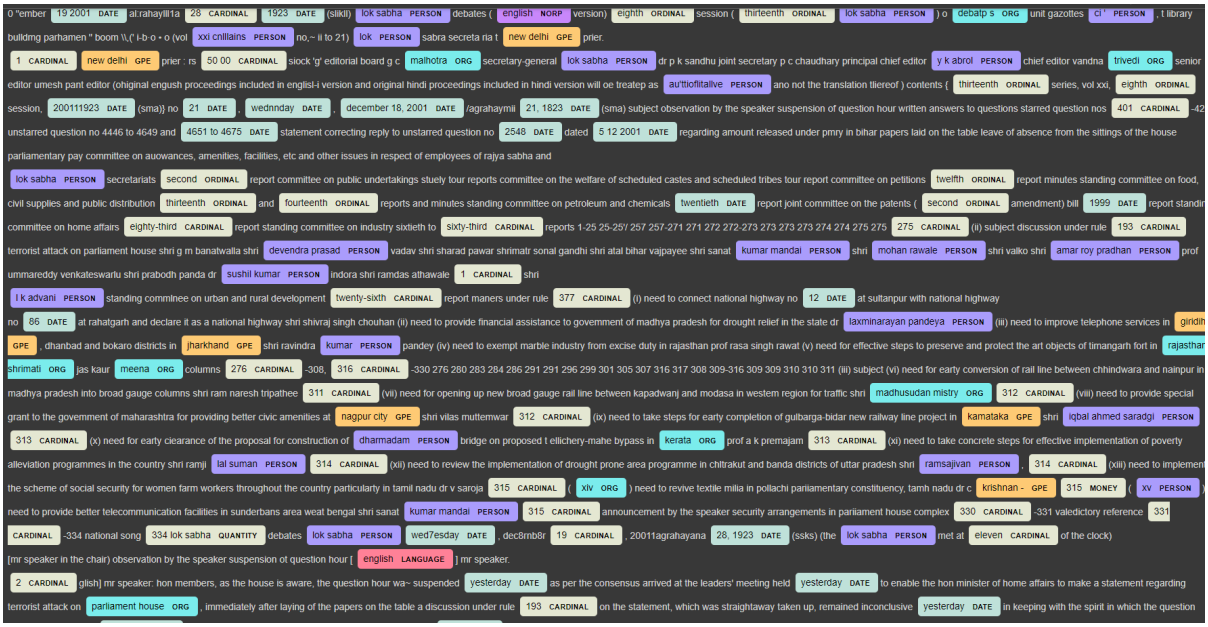
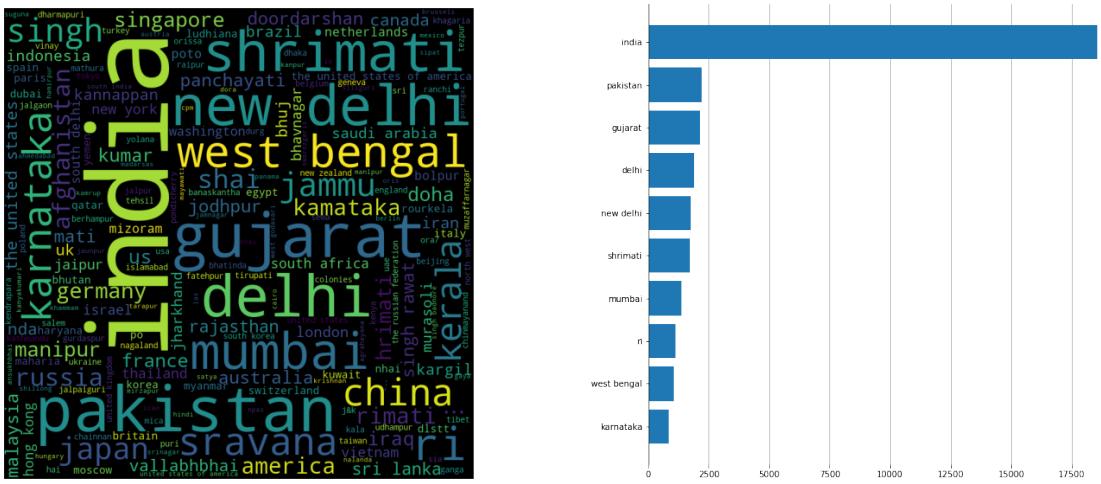


Fig. 3. Text tagged with NER tags, color coded with type of tag i.e. 'PERSON', 'GPE' etc

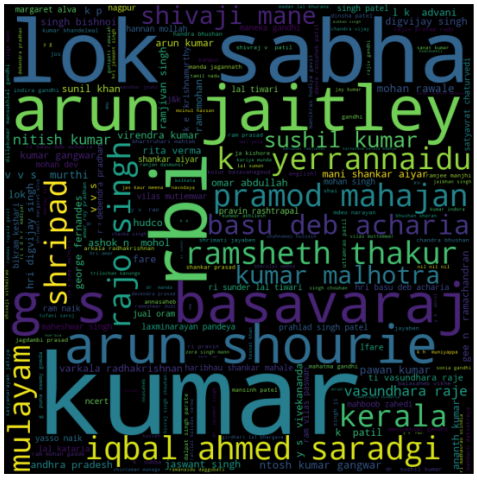


(a) (b)

Fig. 4. GPE tag

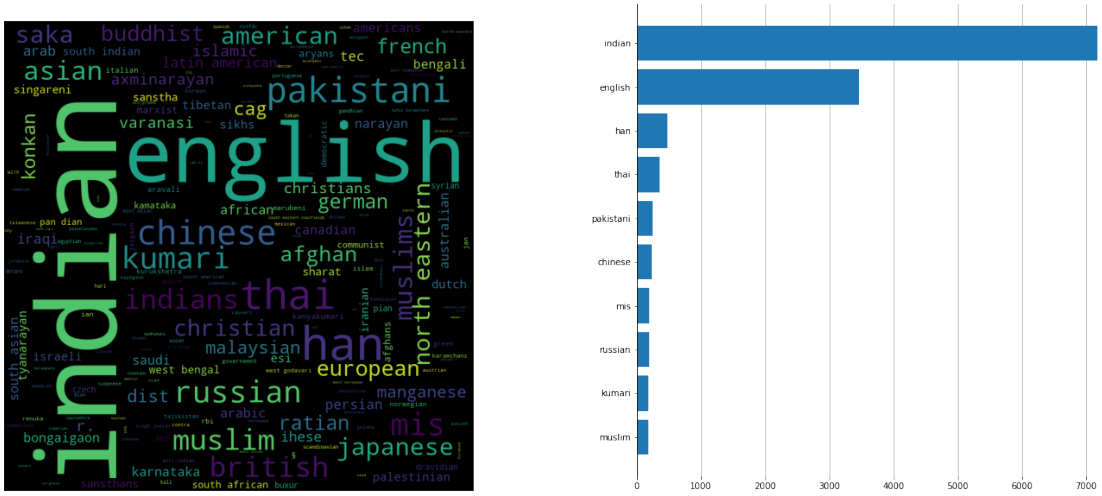
(a) (b)

Fig. 5. LAW tag



(a) (b)

Fig. 6. PERSON tag



(a) (b)

Fig. 7. NORP tag

4.2 DEPENDENCIES

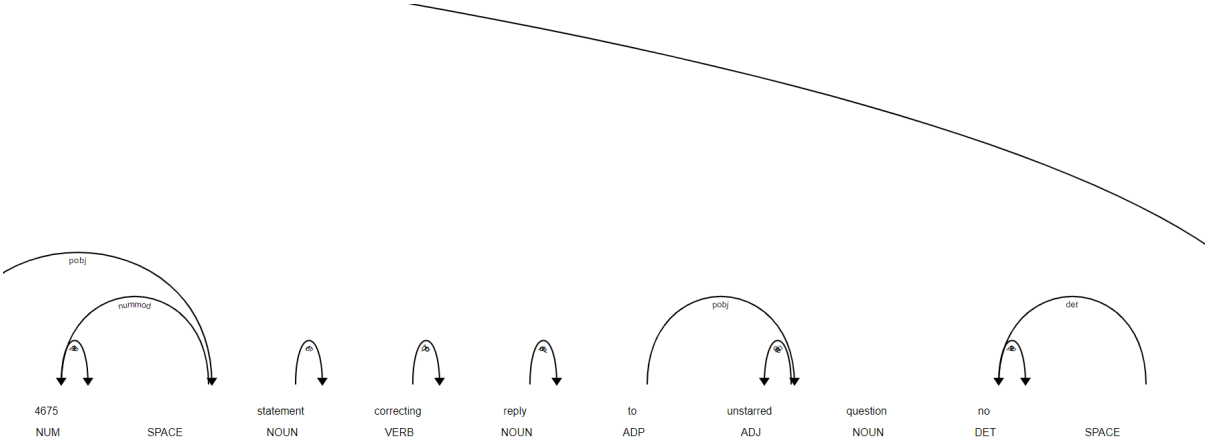


Fig. 8. Depdency parser of 2001 || Link to Video

4.3 INFERENCE

As we can observe from fig. 4, we can see Geographic locations talked most about in speeches were India itself followed by Pakistan, New Delhi, and Gujarat. In addition, we can observe from fig. 5 that section 10(23c) was talked about most. We can analyse the data from this. Also, summary report provides better insights to this.

Entites with high variance such as NORP in fig. 7 are difficult to analyze with bar chart, however in the wordcloud we can see different nationalities that were mentioned during the Lok Sabha debates in year 2001.

From our observation, we can say that results generated by the libraries weren't completely accurate but they still provided relevent data which can be used to analyze the debates over some time-period.

5 CONCLUSION

In conclusion, we developed a chain of systems to help analyze Lok Sabha data on a yearly level. Our web-scrapper downloads and organizes files by year, then summarize and creates a summary from the whole year's Lok Sabha debate data. And NER system generates a ".json" file categorized by entities and each entity having entity name and frequency data for the whole year. This data is then visualised with the help of bar charts and word cloud.

6 CONTRIBUTIONS

Authors and their contributions		
Authors	ID	Contributions
Meet Sable	201901442	pdf2csv, NER tagging, Report
Khushkumar Kantaria	201901299	Scrapper, NER tagging, NER Visualization (Wordcloud and bar charts), Report
Krishna Parmar	201901155	Scrapper, Summarization Script, Dependency Parser, Report

6.1 Reproducibility

Our source code is available on github, containing all descriptions of Jupyter Notebooks made.