IT412 Project Report: Information Processing using Legislative text

AKANKSHA PORWAL, MSC DS, 202118017
DIPSHI JAIN, MSC DS, 202118018, DAIICT, India

The Lok Sabha is known as the House of the Peoples, is the lower house of India's Parliament, with the upper house being the Rajya Sabha. Members of the Lok Sabha are directly elected by the people of India, on the basis of universal suffrage to represent their respective constituencies, and they hold their seats for five years or until the body is dissolved by the President on the advice of the council of ministers.

After the member who initiates discussion on an item of business has spoken, other members can speak on that item of business in such order as the Speaker may call upon them. Only one member can speak at a time and all speeches are directed to the chair. A matter requiring the decision of the House is decided by means of a question put by the Speaker on a motion made by a member. In this project, we have made an attempt to create a dataset of all the Lok Sabha Speeches and perform preliminary analysis on the same. We have created an interface that allows us to understand the locations mentioned during the debates and their frequencies.

Manuscript submitted to ACM

Additional Key Words and Phrases: datasets, Lok Sabha, metaphor detection, sarcasm detection

ACM Reference Format:
AKANKSHA PORWAL, MSC DS, 202118017, DIPSHI JAIN, MSC DS, 202118018. 2018. . In . ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The Lok Sabha is composed of representatives of the people chosen by direct election. The house meets in the Lok Sabha Chambers of the Sansad Bhavan, New Delhi. The maximum strength of the House envisaged by the Constitution is 552. The Constitution empowers the President to summon each House at such intervals that there should not be more than a six-month gap between the two sessions. Hence the Parliament must meet at least twice a year. But, three sessions of Lok Sabha are held in a year. when a motion is putted to the House members for and against to indicate their opinion by saying "Aye" or "No" from their seats. The Chair goes by the voices and declares that the motion is either accepted or rejected by the House.
Three versions of Lok Sabha debates are prepared: the Hindi version, the English version, and the original version. Only the Hindi and English versions are printed. The Hindi version contains proceedings (all questions asked and answers are given thereto and speeches made) in Hindi and verbatim Hindi translation of proceedings in English or in regional languages. The English version contains proceedings in English and the English translation of the proceedings which take place in Hindi or in any regional language.
We have used the printed versions of the debate for our analysis. In the project, we have scrapped the data from those printed debates, preprocessed it, scraped the locations mentioned in each Lok Sabha session and analysed the locations.

2 PROBLEM STATEMENT

The primary goal of this research is to do in-depth analysis of the Lok Sabha debates. To achieve this goal, we divided the entire research into three problems:-

1. The first issue is to scrap the data from the PDFs so that we can easily use that data for the analysis.

2. This section involves labelling the text scraped from PDFs and analysing it.

3. The third issue is to scrap the locations mentioned in the debates of Lok Sabha, validating them and then visualizing it.

3 ABOUT THE DATA

As can be observed from the plot above, the number of hours of work undertaken by Lok Sabha has a downward trend, and the same also gets reflected in the dataset we've created. The amount of data per Lok Sabha in our dataset also has a downward trend, as shown in the figure below

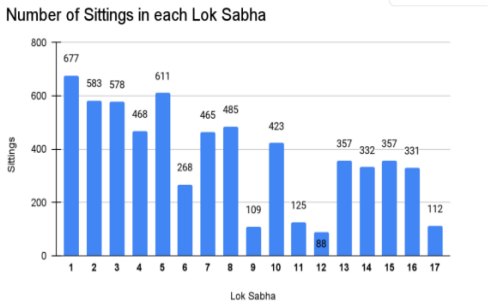


Figure 2: Bar graph showing number of sittings

1.2 Percentage of Hindi Text

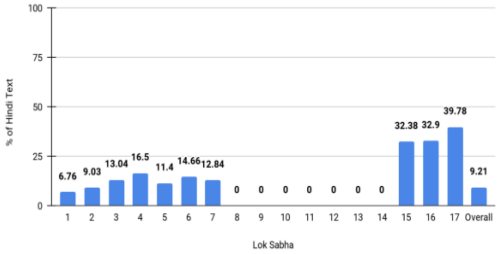
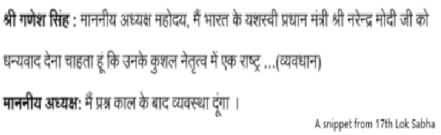


Figure 3: Percentage of Hindi text in each Lok Sabha

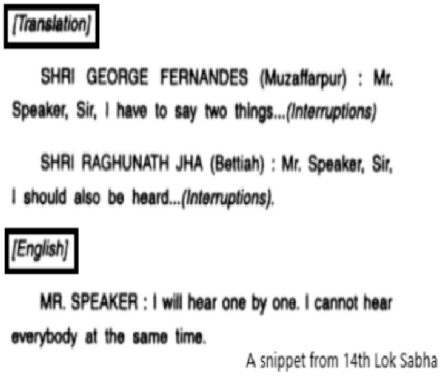
The percentages of Hindi Text for 8th-14th Lok Sabha are 0 because the data contains translated versions of the conversations instead. An example of the same can be seen in the image below.


श्री गणेश सिंह : माननीय अध्यक्ष महोदय, मैं भारत के यशस्वी प्रधान मंत्री श्री नरेन्द्र मोदी जी को धन्यवाद देना चाहता हूं कि उनके कुशल नेतृत्व में एक राष्ट्र ...(व्यवधान)

माननीय अध्यक्ष: मैं प्रश्न काल के बाद व्यवस्था दूंगा ।

A snippet from 17th Lok Sabha

Figure 4: A snippet from 14th Lok Sabha - Hindi


[Translation]

SHRI GEORGE FERNANDES (Muzaffarpur) : Mr. Speaker, Sir, I have to say two things...(Interruptions)

SHRI RAGHUNATH JHA (Bettiah) : Mr. Speaker, Sir, I should also be heard...(Interruptions).

[English]

MR. SPEAKER : I will hear one by one. I cannot hear everybody at the same time.

A snippet from 14th Lok Sabha

4 METHODOLOGY

2.1 Web scraping

• Web scraping was used to obtain raw data.

• This PDF files from the Lok Sabha Digital Library1

2.2 OCR

• The Hindi text encoding was 'WinAnsiEncoding' and the font was 'Aryan2'.

• Since, the 'Aryan2' font was not recognised by the PDF readers (Example: tika2 , pdftotext, PyPDF2), OCR was performed for extracting Hindi text.

• For this purpose, a python library named "ocrmypdf"3 was used.

• OCR was performed for 8 Lok Sabhas, i.e., 3401 documents (1-7, 16 (72 pdfs), 17 Lok Sabhas).

– The rest 8 Lok Sabhas did not have any Hindi text.

2.3 Aryan2 to Unicode Map

• An Aryan2 to Unicode Map was created to help with the recognition of Hindi text.

5 DATA CLEANING AND PROCESSING

• Processed the pdfs obtained after OCR and the ones that were already readable and extracted clean data.

• Saved the extracted clean data in the form of CSVs.

6 NER

• NER was performed using Stanza45, developed by Stanford NLP Group.

• As of 28/10/2021, Lok Sabha NER has been performed on 8th to 14th Lok Sabha since they had only English text.

• A description each label is given in Table 1.

• For Hindi text, we tried NER using Flair and IndicBart. However, due to lack of proper training data and limited RAM, we did not get accurate results.

PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOCATION	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Title of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage (including "%").
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerical that do no fall under another type.

Table 1: Entities recognized by Stanza

7 POS TAGGING

• Lok Sabhas 1 to 7 and 15 to 17 have mixed language (both Hindi and English) data.
• We observed that a speaker may interchange between the two languages frequently.
• To overcome this issue, we run the NER model and the POS tagger on the actual data.
• The NER Model picks up on all English entities (also some Hindi text, but that is wrongly classified), even in the mixed language speeches.
• The POS Tagger picks up on all Hindi data. • We then merge both the models' outputs with accordance to the respective speech.

8 DETAILED ANALYSIS OF ALL THE LOK SABHA DEBATES

3.1 Location Plotting
• In order to get a better understanding of the geographical distribution of the speeches, we plotted all the locations mentioned in all speeches on the world map as seen in Figure 6.
• In the above-mentioned representation, the frequency of the of the locations mentioned was not considered. To understand this, we created a heat-map of the same as seen in Figure 7.
• The following locations were considered: – Entities classified as GPE by English NER.
– For Hindi text, since the NER did not work, we used the "geosky" module in python to get names of all countries, states and cities/towns and translated them in Hindi using Google Translate API. We then performed a lookup of each translated location in the list of entities classified as "PROPN" by Hindi POS-Tagging.
• A total of 8240 locations were plotted (consisting of all the Hindi and English text).

[1]https://github.com/akankshaporwal1205/Lok$_S abha_P roject_7$

9 INTERPRETATION OF OTHER NER PARAMETERS

EVENT. As described in Table 1, EVENT tag is given to historically significant occurrences or natural disasters. When the NER classified text was observed for 16th Lok Sabha, we see the following to events to have to the maximum frequency.
• Independence • the Five Year Plan • the Green Revolution • Olympics Based on our understanding of the Indian Political situation at the time, these topics, other than the Olympics, were a part of the political debate surrounding the new GST bill and the Aadhar Act.

NORP
According to Table 1, NORP tag used for nationalities or religious or political groups. The following were classified as NORP with maximum frequency according to NER.
• Indians • Dalits • Hindu • Tamil • Muslim This set represents the major religions and social groups in India, and most of these have been a part of several political debates for a very long time.

PERSON
PERSON is used for people as seen in Table 1. The following were classified as PERSON with maximum frequency according to NER

LANGUAGE
Any named language is classified as LANGUAGE by NER as observed in Table 1. The following were classified as LANGUAGE with maximum frequency according to NER.
• English
• Hindi
• Tamil
• Marathi
• Telugu

1

10 CONTRIBUTIONS

AKANKSHA PORWAL 202118017

- Hindi NER/POS tagging
- English NER/ POS tagging
- Visualisation

DIPSHI JAIN 202118018

- Geocoding nominatim
- Calculate stats
- Visualisation