Constitution of India: Transformer-based NLP Analysis

-Vikas Tunge 2022OCTVUGP0031)

1. Problem Statement

The objective of this project is to analyze the Constitution of India using **only transformer-based NLP pipelines** in order to extract meaningful insights about its structure, themes, and institutional focus. The task requires converting an unstructured legal document into structured knowledge, supported by visualizations and a coherent narrative, while also reporting appropriate evaluation metrics.

2. Overall Approach

The Constitution is a long, complex, and unlabeled legal document. Traditional NLP methods are insufficient due to long-range dependencies and domain-specific language. Therefore, we adopt **pretrained transformer pipelines** to:

- Segment the document into meaningful units
- Identify named entities
- Discover dominant themes
- Classify sentences into functional categories
- Aggregate findings into interpretable visual insights

Each notebook contributes one logical layer to this pipeline.

3. Methodology and Analysis Stages

3.1 Text Extraction & Sentence Segmentation

What was done:

- Extracted raw text from the official Constitution PDF
- Cleaned formatting noise introduced by PDF layout
- Segmented the document into sentence-level units

Why it was done:
Transformers have input length limits and operate best on semantically complete units. Sentence-level segmentation ensures stable, scalable processing for downstream tasks like NER and classification.

Outcome:
A clean corpus of constitutional sentences suitable for transformer pipelines.

3.2 Named Entity Recognition (NER)

What was done:

- Applied a pretrained transformer-based NER model
- Extracted entities such as organizations, locations, and persons
- Aggregated entity frequencies and distributions

Why it was done:
NER reveals the **institutional and administrative focus** of the Constitution by identifying which entities are emphasized throughout the text.

Outcome:
Clear evidence of institutional dominance (e.g., Parliament, President, States), supporting the narrative that the Constitution is governance-centric rather than individual-centric.

3.3 Topic Modeling via Embeddings

What was done:

- Generated dense sentence embeddings using transformer encoders
- Applied clustering to group semantically similar sentences

Why it was done:
Legal documents contain latent themes not explicitly labeled. Embedding-based topic discovery uncovers these themes without manual annotation.

Outcome:
Identification of major constitutional themes such as fundamental rights, governance structure, judiciary, and federal relations.

3.4 Sentence Classification

What was done:

- Used zero-shot transformer classification
- Assigned sentences to predefined legal-functional categories (e.g., Rights, Duties, Powers, Procedures)

Why it was done:
This allows functional interpretation of the Constitution, showing *how* text operates legally rather than just *what* it discusses.

Outcome:
Quantified functional distribution of constitutional text, enabling comparison between rights-oriented and authority-oriented provisions.

3.5 Dashboard & Combined Visualizations

What was done:

- Integrated outputs from all previous notebooks
- Created charts summarizing entities, topics, and classifications

Why it was done:
Visual aggregation transforms raw NLP outputs into interpretable insights and supports a coherent analytical narrative.

Outcome:
A unified view of constitutional structure and emphasis, suitable for academic evaluation and presentation.

4. Concluding Summary

This project successfully applies transformer-based NLP to convert the Constitution of India from unstructured legal text into structured analytical insights, meeting all technical and interpretive requirements of the problem statement.