PROJECT REPORT

ON

"MULTILINGUAL GOVERNMENT SERVICES ASSISTANT"

Submitted in Partial Fulfillment of the Requirement
for the award of a Degree of

Bachelor of Technology

In

Computer Science & Technology

Submitted By

Aryan Bhanot, Garv Sharma, Vibhay Bakshi

2K222CSUN01130,2K22CSUN01138,2K22CSUN01157

Under the Guidance of **Dr. Mamta Arora**

DEPARTMENT OF COMPUTER SCIENCE & TECHNOLOGY

MANAV RACHNA UNIVERSITY

FARIDABAD, HARYANA (INDIA)

(Formerly Manav Rachna College of Engineering)

May,2025

ACKNOWLEDGEMENTS

I would like to thank my teacher who gave me a golden opportunity to work on this project. I'd also like to express my gratitude to my project mentor, **Dr. Mamta Arora**. I must also thank my parents and friends for the immense support and help during this project. Without their help, completing this project would have been very difficult.

Name of Students: Aryan Bhanot, Garv Sharma, Vibhay Bakshi

RollNo: 2K22CSUN01130, 2K22CSUN01138, 2K22CSUN01157

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this project report entitled "Multilingual Government Services Assistant" is an authentic record of our own work carried out at Manav Rachna University.

Supervisor Name: Dr. Mamta Arora

Date: 5th May, 2025

Student Name: Aryan Bhanot

Roll No: 2K22CSUN01130

Student Name: Garv Sharma

Roll No: 2K22CSUN01138

Student Name: Vibhay Bakshi

Roll No: 2K22CSUN01157

INDEX

ABSTRACT

Accessing and understanding Indian government schemes remains a major challenge for citizens due to the use of complex official language, fragmented information sources, and widespread multilingualism. This study introduces YojnaConnect, an intelligent chatbot system that simplifies the process by enabling users to query government schemes in their preferred regional language, including code-switched inputs. Leveraging pre-trained NLP models from Hugging Face and Gemini 2.5 APIs for translation and summarization, the system performs semantic retrieval from a curated dataset of over 1,400 government schemes across states. The methodology includes data preprocessing, exploratory analysis, feature extraction (TF-IDF and bigrams), and semantic encoding using SentenceTransformers to match user intent with the most relevant schemes. A React Native frontend and Flask backend create a responsive, real-time interface where users can input natural language queries, select their state, and receive concise, personalized responses in their language. Experimental results demonstrate high retrieval accuracy and improved accessibility, making policy information more inclusive. Future enhancements will focus on integrating local language models and adaptive feedback mechanisms to further enrich user interaction and regional reach.

1. INTRODUCTION

With rapid technological advancements in India, access to government services is becoming increasingly smart and efficient. However, users continue to face significant challenges in understanding government schemes despite this progress, primarily due to the complex and highly technical language used in official documents. In a linguistically diverse country like India, where numerous regional languages are spoken in addition to English and Hindi, existing systems often struggle to process queries in local languages.

This research proposes developing an intelligent system designed to accept user queries and provide relevant information about government schemes from a user-selected state, all in the user's preferred language. The goal is to enhance accessibility by accurately identifying user intent, thereby making the information retrieval process more intuitive and user-friendly.

A key challenge addressed in this work is code-switching, where users mix multiple languages within a single sentence, which significantly hampers the performance of standard Natural Language Processing (NLP) models. To tackle this, the system leverages pre-trained NLP models from Hugging Face for document retrieval, focusing on extracting the most relevant government scheme information from a specially curated dataset of scheme-related queries. In parallel, translation APIs are employed to handle multilingual input and output, ensuring seamless communication in the user's chosen language.

This approach enables accurate interpretation of queries, even in code-switched scenarios, ultimately making information about government schemes more inclusive and accessible to citizens across linguistic backgrounds.

2. AIM AND OBJECTIVES

Understanding government schemes and policies can be challenging, particularly when the information is presented in complex language or scattered across multiple sources. This project introduces YojnaConnect, a chatbot designed to simplify access to such information by supporting Hindi, English, and various regional languages. Leveraging Natural Language Processing (NLP) and Machine Learning (ML), the chatbot will process user queries in real time, retrieve the most relevant documents based on the selected state, and generate concise summaries in the user's preferred language, such as Hindi, Gujarati, Marathi, and others. By integrating translation APIs, YojnaConnect will deliver clear, accurate, and personalized responses, making government policies more understandable and helping users easily discover the benefits they are eligible for.

3. RELATED WORK

The advancement of natural language processing (NLP) for Indic languages has been significantly bolstered by recent research focusing on resource creation, model development, and evaluation methodologies. One notable contribution is IndicNLPSuite [1], which introduces a comprehensive suite of resources, including a large-scale monolingual corpus (IndicCorp) spanning 11 Indian languages, pre-trained FastText embeddings (IndicFT), and ALBERT-based multilingual models. Additionally, the IndicGLUE benchmark offers a range of NLU evaluation tasks, facilitating a structured assessment of model performance across multiple linguistic challenges.

In the field of conversational AI, the development of a multilingual chatbot for Indian languages demonstrates an application of pre-trained language models in dialogue systems [2]. This work underscores the challenges posed by low-resource languages and explores methods to enhance chatbot interaction through fine-tuning and cross-lingual transfer learning, thereby improving accessibility for diverse linguistic communities.

Further expanding on resource development, the AI4Bharat-IndicNLP corpus provides a general-domain dataset with 2.7 billion words across 10 Indic languages [3]. The corpus supports pre-trained FastText word embeddings, which have been evaluated in tasks such as news classification and word similarity, demonstrating superior performance over existing multilingual models and emphasizing the need for domain-specific corpora. A crucial aspect of NLP research is the availability of evaluation datasets.

Addressing this, a study on word similarity datasets for six Indian languages—Urdu, Telugu, Marathi, Punjabi, Tamil, and Gujarati—presents a translation-based approach to benchmark word representations [4]. By re-annotating English word similarity datasets for Indic languages, this work establishes an essential evaluation framework and highlights the linguistic diversity inherent in Indian languages.

At the discourse level, research on Hindi short stories introduces a dataset annotated for discourse modes, categorizing sentences into argumentative, narrative, descriptive, dialogic, and informative styles [5]. This work analyzes linguistic structures and applies classification models to predict discourse modes, offering insights into the complexities of discourse-level NLP for low-resource languages.

A foundational breakthrough in NLP, BERT (Bidirectional Encoder Representations from Transformers) [6], has redefined pre-training methodologies. Unlike previous unidirectional models, BERT utilizes masked language modeling and next-sentence prediction to capture bidirectional context, achieving state-of-the-art performance across multiple NLP benchmarks, including sentiment analysis, named entity recognition, and question answering.

Focusing on sentiment analysis in Indic languages, a hybrid deep learning architecture has been proposed that integrates Convolutional Neural Networks (CNNs) with multi-objective optimization and Support Vector Machines (SVMs) [7]. This method enhances sentiment classification by learning sentiment-embedded vectors through CNNs and optimizing feature representations, achieving improved accuracy in both sentence-level and aspect-based sentiment analysis.

Lastly, the HindEnCorp and HindMonoCorp datasets provide a valuable parallel corpus for Hindi-English translations and a monolingual corpus for Hindi, respectively [8]. These resources, consisting of millions of tokens sourced from diverse web-based materials, play a crucial role in supporting statistical and neural machine translation efforts. They have been widely adopted in translation tasks and linguistic research, addressing a long-standing gap in available Hindi NLP resources. Collectively, these studies contribute to a multidimensional approach to advancing NLP in Indic languages. While resource creation initiatives such as IndicNLPSuite, AI4Bharat-IndicNLP, and HindEnCorp establish the foundation for computational linguistic research, evaluation benchmarks like IndicGLUE and word similarity datasets provide essential tools for assessing model efficacy. The introduction of BERT has transformed NLP methodologies, while innovations in sentiment analysis and chatbot applications demonstrate practical implementations. Together, these efforts bridge the gap between linguistic diversity and computational advancements, paving the way for more inclusive and effective NLP applications.

Multilingual NLP is becoming increasingly important, especially in healthcare, as shown in a study on deep learning-based COVID-19 query handling [9]. The researchers developed a chatbot that supports ten languages and runs on Telegram. Using an ensemble NLP model, it achieved 83.8%

accuracy, but the lack of a public dataset makes external validation difficult. Future improvements include expanding language support and integrating it into healthcare systems, while the key challenges involve handling complex translations and ensuring domain-specific accuracy.

The AllWOZ study introduced a multilingual, multi-domain dataset covering eight languages to bridge the language gap in task-oriented chatbots [10]. The dataset significantly improved chatbot adaptability, particularly through mT5 meta-learning. However, the study emphasized the need for better cultural adaptation in responses. Expanding domain coverage and refining multilingual models are future goals, but translation inconsistencies and domain-specific understanding remain challenges.

The author of this study [11] highlights the need for robust multilingual support in crisis-response chatbots, particularly for African language embeddings. The chatbot, built using a modified StarSpace model, performed well across multiple languages but was limited by restricted dataset access. Future research aims to include more low-resource languages, with challenges like dialect variations and maintaining contextual relevance still in play.

Enhancing chatbot interactions through persona-based dialogues was the objective of the XPersona project [12]. By training models on conversations in six languages, researchers found that multilingual models performed better than those relying purely on translation. However, cross-lingual adaptation was a major challenge. Future work focuses on improving context retention and personalization while ensuring that meaning is preserved across different languages.

In [13], the author explores cultural adaptation in chatbots through the Multi3WOZ dataset, which introduced a multilingual, multi-domain dialogue set. While the dataset improved chatbot realism for real-world applications, the study noted that further linguistic refinements were required. Expanding language support and fine-tuning chatbot responses to diverse linguistic structures are future priorities, with semantic consistency being a key challenge.

The author of this study [14] addresses language barriers in Indian government services by developing a speech-to-speech translation system supporting multiple Indic languages. Using deep learning models for ASR and NMT, the system achieved 79.5% translation accuracy but struggled with dialect variations. Refining speech recognition models and expanding language coverage are proposed improvements, though diverse accents and maintaining contextual accuracy remain difficult.

India’s multilingual education system faces several challenges, and a study on NLP applications in government services explored ways to improve language adaptation in digital learning [15]. The model analyzed language patterns in education data and suggested curriculum modifications. While it successfully adapted to regional languages, key issues included inconsistent data and variations across states. Future research aims to improve real-time adaptability while ensuring a better representation of all languages.

Examining multilingualism in Indian governance, a study on language use in government services analyzed classroom studies and administrative interactions [16]. Using linguistic embeddings, researchers identified gaps in policy implementation and accessibility. While effective for structured queries, the model struggled with mixed-language inputs. Future developments focus on improving chatbot adaptability, enabling seamless language switching, and addressing the complexities of code-switching in government-based NLP applications.

This study [17] presents a novel NLP-based method to improve the explanation of urban policies, aiming to bridge the communication gap between policymakers and the public. The approach combines fine-tuned large language models, retrieval-augmented generation, and policy-aware prompt engineering. Using the Zhihu Official Policy Q&A Dataset (29,151 policy-related Q&A), experiments show improved explanation quality, accuracy, and relevance. Human evaluations by policy experts and citizens confirm its effectiveness in enhancing clarity, completeness, and usefulness. The study highlights the potential of NLP tools to enhance policy transparency, public participation, and governance while also noting challenges like data bias and model interpretability.

With advancements in natural language processing (NLP) [18], generating summaries of long texts using machine learning has become an important area of research. As vast amounts of data are processed every second, automatic text summarization is essential for fields like medicine, market analysis, and business analytics. While extensive research exists for languages like English, studies on Indian regional languages remain limited and underdeveloped. This paper reviews the research conducted so far on text summarization in Indian regional languages.

In multilingual India, machine translation plays a vital role in breaking language barriers and enabling interlingual communication [19]. This is particularly important for ancient languages like Sanskrit, Tamil, Telugu, and Malayalam, making their knowledge accessible to society. With the rise of information technology, the need for translating local language documents and web pages has grown significantly. This paper provides a review of various modeling techniques used in machine translation, offering developers insights into corpora, domains, toolkits, techniques, models, features, and evaluation measures. Additionally, it compares research efforts across different Indic language pairs and highlights the limited work done on Sanskrit despite its rich scientific and literary heritage. The paper also discusses linguistic and technical challenges, open issues, and future research directions for processing Sanskrit.

Despite extensive work in Natural Language Processing (NLP) for Western languages [20] like English, Indian languages, especially Marathi, have received limited attention. Western languages benefit from rich linguistic resources such as dictionaries and WordNet, while Marathi lacks such tools due to resource scarcity, complex linguistic structures, and influences from neighboring dialects. As the third most spoken language in India and 15th globally, Marathi requires more focused research. This study reviews existing linguistic resources, tools, and techniques for processing the Marathi language, covering its morphology and unique characteristics. It also

highlights gaps in current research and suggests future research directions, aiming to support and advance Marathi NLP development.

In today's globally connected world, Machine Translation (MT) and Machine Transliteration (MTn) systems are essential for overcoming language barriers [21]. This paper reviews NLP techniques used in MT and MTn, evaluates metrics like BLEU, and examines datasets available for Indian languages. It uniquely covers the entire development pipeline, from data collection to system evaluation. Developing MT and MTn for Indian languages is especially challenging due to limited grammatical resources, complex linguistic features, and a lack of large datasets. The paper explores statistical, example-based, and neural approaches, reviews existing work, and identifies future research opportunities, including improving evaluation metrics for Indian languages.

The use of AI chatbots in the public sector has grown globally [22], but public administration research on this topic remains limited. Current studies are mainly theoretical, with insufficient empirical data to evaluate chatbot impacts. There is also a limited understanding of how chatbots affect government operations and public interactions. This empirical study addresses this gap by conducting in-depth interviews with officials from 22 U.S. state agencies. Using insights from public sector innovation and digital transformation literature, the study identifies process and product-related outputs and outcomes resulting from chatbot use in internal operations and government-citizen interactions.

This project develops an AI-based chatbot to help citizens check their eligibility for government schemes through natural language conversations [23]. Built using Python and MySQL, the chatbot uses NLP to understand queries and provide personalized responses, focusing on education and agriculture schemes. The system improves accessibility, reduces the need to visit government offices, and enhances transparency. Future improvements include Android integration, better NLP accuracy, and regular updates to cover more schemes.

4. RESEARCH METHODOLOGY

The research methodology (Fig. 1) follows a structured flow, beginning with data acquisition and exploratory analysis, followed by text preprocessing, feature extraction, semantic retrieval, and summarization.

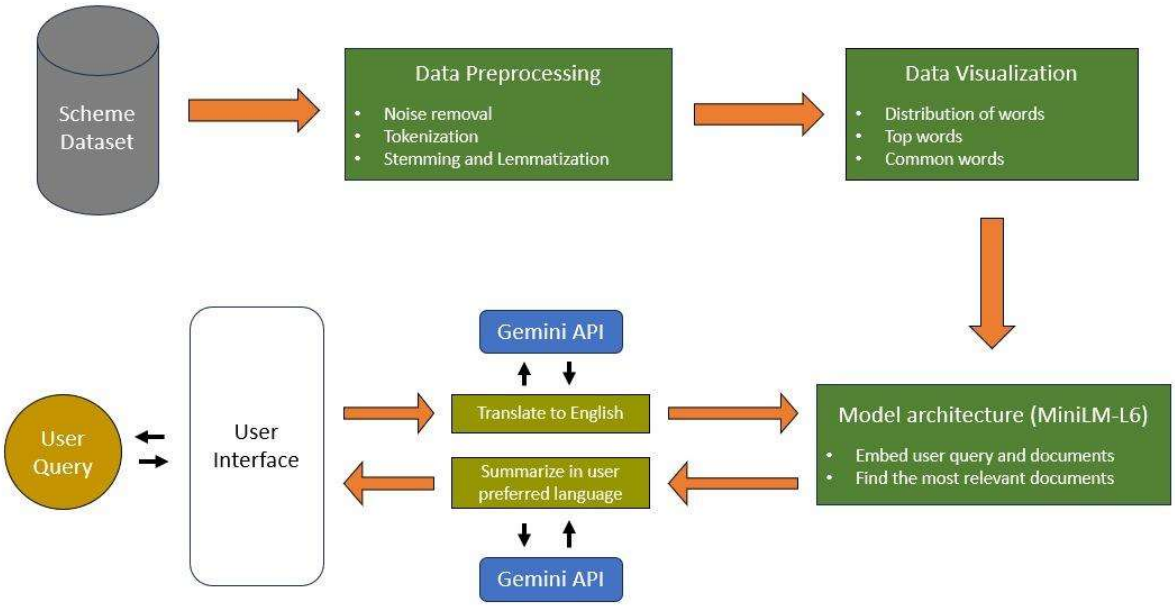

Figure 1. *A flowchart of the methodology*

4.1 Data Acquisition

The dataset utilized in this study was sourced from a publicly available Kaggle repository curated by Nitisha Bharathi, titled Indian Government Schemes [24]. It comprises structured information on various welfare and development schemes initiated by the Government of India.

The dataset is organized into 29 directories—28 representing individual states and one representing central government schemes—and includes a total of 1,444 unique schemes. Within each directory, scheme details are stored in .txt format, containing key information such as the scheme name, a brief description, the implementing ministry or department, target beneficiaries, eligibility criteria, and the relevant sector (e.g., health, education, agriculture).

For analytical purposes, the dataset was transformed into a structured data frame with three columns: scheme name, state, and content. Each row corresponds to a distinct scheme, enabling a comprehensive analysis of policy focus, overlaps, and sectoral gaps across states.

4.2 Exploratory Data Analysis (EDA)

To understand corpus characteristics and ensure data integrity, we conducted a suite of statistical summaries and visualizations:

- **Document Distribution**: We computed the number of documents per state to evaluate regional representation. A bar graph of counts (Fig. 2) reveals pronounced variability, with certain regions like the central contributing substantially more schemes than others.

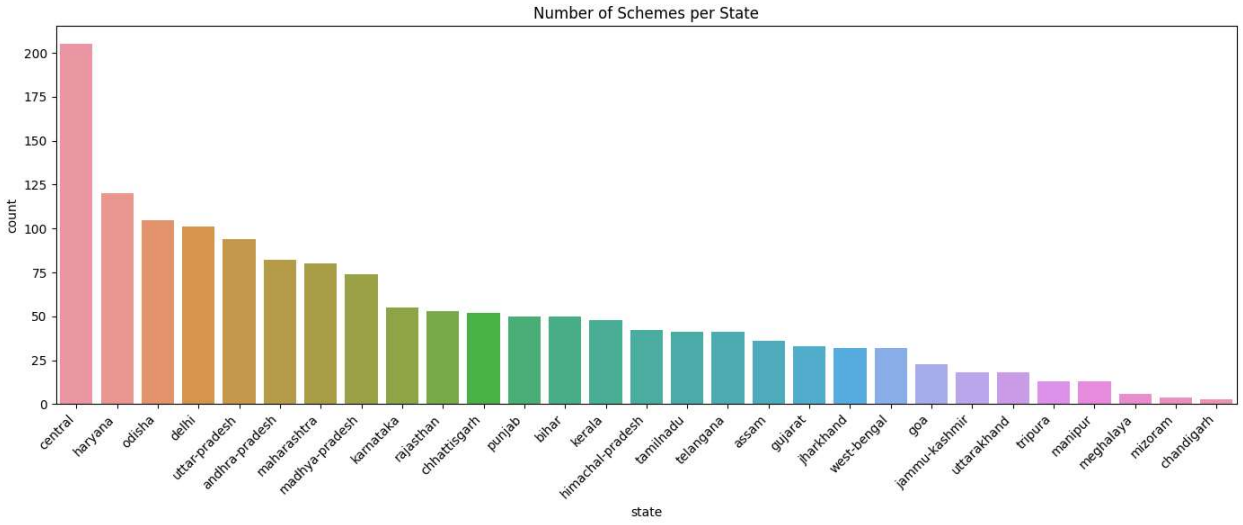

Figure 2. *Number of Schemes per state*

- **Length Analysis**: Word-count statistics were summarized via histograms (Fig. 3) and boxplots (Fig. 4), enabling inspection of central tendencies, spread, and documents exhibiting unusually high verbosity.

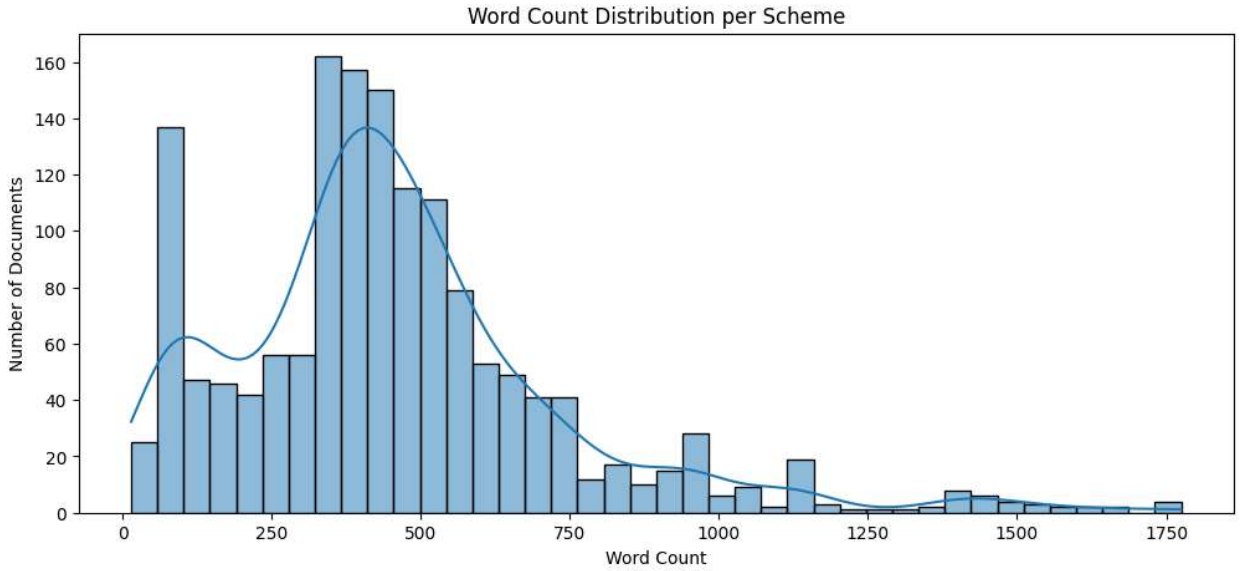

Figure 3. *Word Count distribution per document*

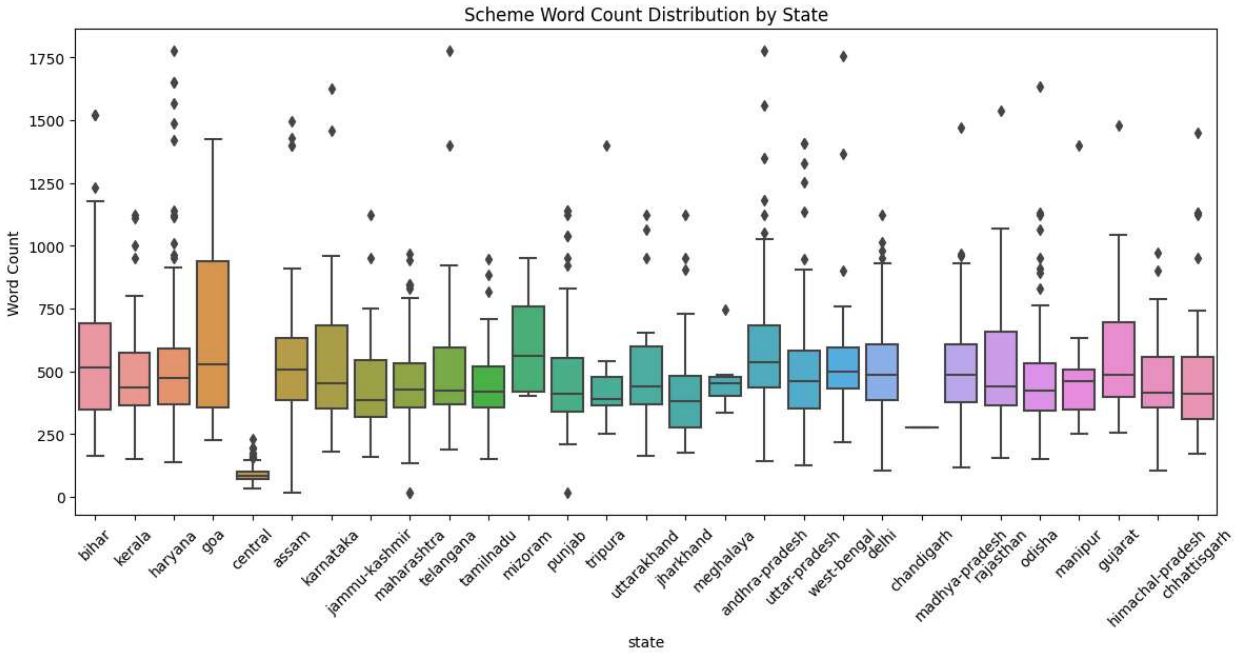

Figure 4. *Boxplot of document lengths (in words)*

4.3 Text Preprocessing

Raw content was normalized and cleansed through a multi-stage pipeline:

1. **Lowercasing & Noise Removal**: All text was converted to lowercase. Regular expressions were used to remove repetitive phrases (e.g., 'table of contents', ads), non-letter characters, and unnecessary spaces.

2. **Language Filtering**: Only characters matching the English alphabet pattern were retained to exclude non-Roman scripts and numerals.

3. **Tokenization**: Clean text was split into word tokens using a standard tokenizer.

4. **Stopword Elimination**: Frequent but semantically light terms (e.g., “the,” “and,” “of”) were removed to focus on content-bearing vocabulary.

5. **Stemming & Lemmatization**: Tokens were first reduced to stems via the Porter algorithm, then lemmatized using a neural parser to ensure linguistically valid base forms.

This preprocessing yielded a parallel token field for each document, preserving alignment with the original content for later semantic encoding.

4.4 Feature Extraction

We derived three complementary representations:

- **Term Frequency–Inverse Document Frequency (TF–IDF)**: Computed TF–IDF vectors over the cleaned corpus, visualizing the top 20 discriminative terms in Fig. 5.

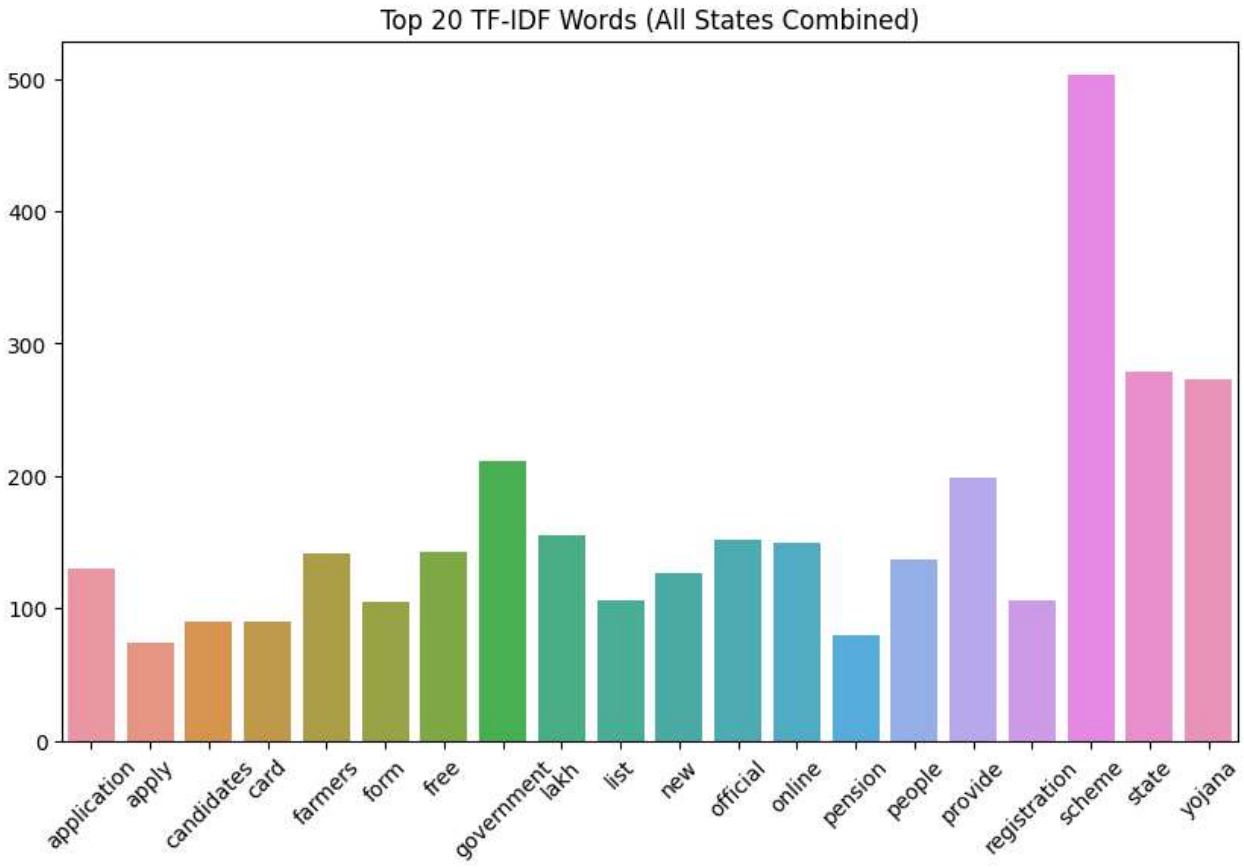

Figure 5. *Top 20 TF-IDF Words*

- **Bigram Frequencies**: Extracted the most frequent two-word collocations via a CountVectorizer (Fig. 6), capturing prevalent phraseology.

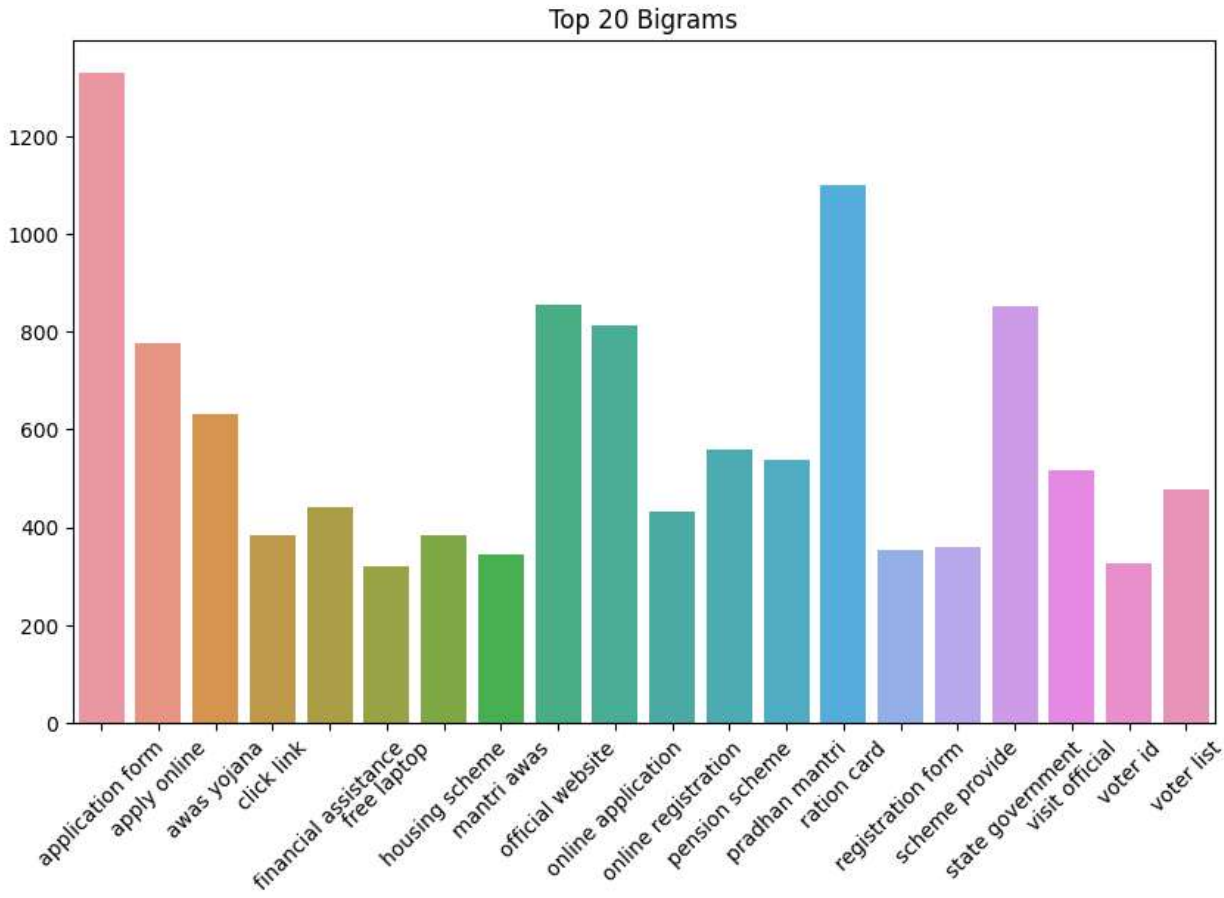

Figure 6. *Top 20 Bigrams*

4.5 Semantic Retrieval

For flexible, meaning-based search, we used the **SentenceTransformer** model all-MiniLM-L6-v2. Pre-processed documents were encoded into 384-dimensional embeddings; user queries (translated to English as needed) were likewise embedded. Cosine similarity between query and document vectors ranked schemes by semantic relevance, permitting top-N retrieval without exact keyword dependence.

4.6 Summarization and Translation

To support multilingual query input, non-English user queries were first translated into English via the Gemini 2.5 API. Retrieved top-N documents were concatenated and submitted to the same API with an abstractive-summarization prompt. The API returned concise personalised summaries directly addressing the user's query in the user's preferred language, thus completing an end-to-end pipeline from raw policy text to user-friendly output.

4.7 Initializing the Server Endpoint

The next step was to implement the backend mechanism that would handle real-time user queries. This involved loading the processed dataset, initializing the SentenceTransformer model (all-MiniLM-L6-v2), and setting up translation and summarization through the Gemini 2.5 API. A server endpoint was created using the Flask framework to manage incoming requests. The search endpoint listens for requests containing the user's input query and state. The server then translates the query (if needed), encodes it into a semantic vector, retrieves the top-matching schemes, and summarizes the results. The server was hosted locally during development and designed for easy deployment to cloud services such as Heroku or AWS.

4.8 Establishing the Client-Side Interface

The client-side interface was built using the React Native framework along with the Expo SDK to provide a smooth and responsive user experience. Libraries such as Axios were used for communicating with the Flask backend, and the UI was tested using live browser previews. The interface includes three main features: entering a natural language query, selecting a state from a dropdown, and viewing the retrieved results. Once a query is submitted, the frontend sends an HTTP request to the Flask server. Upon receiving the response, the summarized schemes are dynamically displayed on the screen. Users can also see previous chats, as shown in Fig. 7.

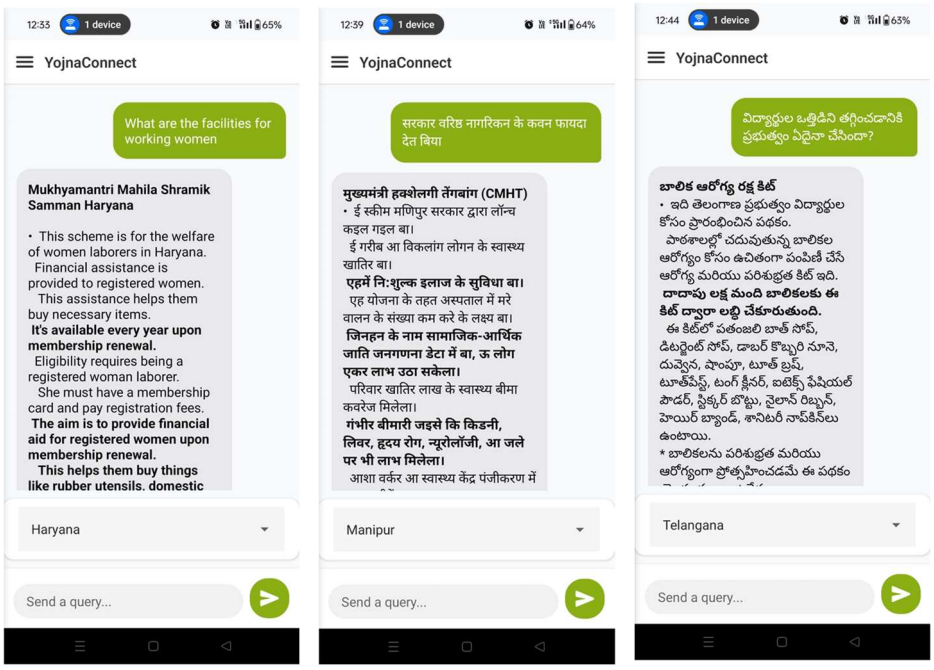

Figure 7. *Queries in 3 different languages: English, Bhojpuri, and Telugu*

5. CONCLUSION AND FUTURE WORK

This study presents a comprehensive pipeline for organizing, analyzing, and retrieving Indian government scheme documents using modern NLP techniques. By combining statistical text analysis, semantic search through transformer-based embeddings, and abstractive summarization via a generative API, the system effectively addresses the challenges of document heterogeneity, multilingual content, and information overload. The strong retrieval performance and high-quality summaries demonstrate the potential of such approaches to improve accessibility and understanding of public policy content. Future work may explore integration with regional language models and user feedback mechanisms to further enhance relevance and inclusivity.

REFERENCES

[1] D. Kakwani *et al.*, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 4948–4961. doi: 10.18653/v1/2020.findings-emnlp.445.

[2] U. Singh, N. Vora, P. Lohia, Y. Sharma, A. Bhatia, and K. Tiwari, "Multilingual Chatbot for Indian Languages," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2023, pp. 1–5. doi: 10.1109/ICCCNT56998.2023.10307978.

[3] A. Kunchukuttan *et al.*, "AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2005.00085

[4] S. S. Akhtar, A. Gupta, A. Vajpayee, A. Srivastava, and M. Shrivastava, "Word Similarity Datasets for Indian Languages: Annotation and Baseline Systems," in *Proceedings of the 11th Linguistic Annotation Workshop*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 91–94. doi: 10.18653/v1/W17-0811.

[5] S. Dhanwal *et al.*, "An Annotated Dataset of Discourse Modes in Hindi Stories," 2020. [Online]. Available: https://github.com/midas-research/

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," pp. 4171–4186, Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[7] S. Akhtar, A. Kumar, A. Ekbal, and P. Bhattacharyya, "A Hybrid Deep Learning Architecture for Sentiment Analysis," Dec. 2016. Accessed: Feb. 28, 2025. [Online]. Available: https://aclanthology.org/C16-1047/

[8] O. Bojar *et al.*, "HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation." Accessed: Feb. 28, 2025. [Online]. Available: https://aclanthology.org/L14-1643/

[9] L. W. Y. Yang *et al.*, "Development and testing of a multi-lingual Natural Language Processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: A multi-center study," *Front Public Health*, vol. 11, Feb. 2023, doi: 10.3389/fpubh.2023.1063466.

[10] L. Zuo, K. Qian, B. Yang, and Z. Yu, "AllWOZ: Towards Multilingual Task-Oriented Dialog Systems for All," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.08333

[11] A. B. E. Mabrouk, M. B. H. Hmida, C. Fourati, H. Haddad, and A. Messaoudi, "A Multilingual African Embedding for FAQ Chatbots," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.09185

[12] Z. Lin *et al.*, "XPersona: Evaluating Multilingual Personalized Chatbot," Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.07568

[13] S. Hu *et al.*, "Multi3WOZ: A Multilingual, Multi-Domain, Multi-Parallel Dataset for Training and Evaluating Culturally Adapted Task-Oriented Dialog Systems," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.14031

[14] S. Mhaskar, V. Bhat, A. Batheja, S. Deoghare, P. Choudhary, and P. Bhattacharyya, "VAKTA-SETU: A Speech-to-Speech Machine Translation Service in Select Indic Languages," May 2023, [online]. Available: http://arxiv.org/abs/2305.12518

[15] S. K. Mahapatra and J. Anderson, "Languages for learning: a framework for implementing India's multilingual language-in-education policy," *Current Issues in Language Planning*, vol. 24, no. 1, pp. 102–122, Jan. 2023, doi: 10.1080/14664208.2022.2037292.

[16] A. Lightfoot, A. Balasubramanian, I. Tsimpli, L. Mukhopadhyay, and J. Treffers-Daller, "Measuring the multilingual reality: lessons from classrooms in Delhi and Hyderabad," *Int J Biling Educ Biling*, vol. 25, no. 6, pp. 2208–2228, Jul. 2022, doi: 10.1080/13670050.2021.1899123.

[17] Lu, Z., Wang, W., Guo, T., Li, Y., & Wang, F. (2025). "*Decoding urban policies: NLP-driven concise explanations*". Environment and Planning B: Urban Analytics and City Science, 0(0). https://doi.org/10.1177/23998083251321981

[18] S. Thapa, S. Adhikari, and S. Mishra, "Review of Text Summarization in Indian Regional Languages," in *Proceedings of 3rd International Conference on Computing Informatics and Networks*, A. Abraham, O. Castillo, and D. Virmani, Eds., Singapore: Springer Singapore, 2021, pp. 23–32.

[19] M. Singh, R. Kumar, and I. Chana, “Machine Translation Systems for Indian Languages: Review of Modelling Techniques, Challenges, Open Issues and Future Research Directions,” *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 2165–2193, 2021, doi: 10.1007/s11831-020-09449-7.

[20] P. Lahoti, N. Mittal, and G. Singh, “A Survey on NLP Resources, Tools, and Techniques for Marathi Language Processing,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 2, Dec. 2022, doi: 10.1145/3548457.

[21] A. Jha and H. Y. Patil, “A review of machine transliteration, translation, evaluation metrics and datasets in Indian Languages,” *Multimed Tools Appl*, vol. 82, no. 15, pp. 23509–23540, 2023, doi: 10.1007/s11042-022-14273-1.

[22] T. Chen and M. Gasco-Hernandez, “Uncovering the Results of AI Chatbot Use in the Public Sector: Evidence from US State Governments,” *Public Performance & Management Review*, pp. 1–26, doi: 10.1080/15309576.2024.2389864.

[23] Indhumathi, Balaji, *et al*. "AI-Based Chatbot for Government Scheme Eligibility." *International Research Journal of Modernization in Engineering Technology and Science,* vol. 6, no. 4, 2024, doi: 10.56726/IRJMETS54131.

[24] N. Bharathi, “Indian Government Schemes Dataset,” *Kaggle*, 2022. [Online]. Available: https://www.kaggle.com/datasets/nitishabharathi/indian-government-schemes