Information Retrieval Project on Public Opinion Analysis for Policy Makers

Jayshil Shah[1], Kirtan Shah[2], Rohit Gupta[3], Ankur Bhoi[4], Amit Kumar Prajapati[5], Gaurav Deswal[6]
Indraprastha Institute of Information Technology, Delhi, India
{jayshil23138, kirtan23086, rohit23078, ankur23112, amit23110, gaurav23118}@iiitd.ac.in

***Abstract*—This report presents a novel approach to analyze public opinion on government policies using Reddit comments. Users select a policy via a dropdown menu, and comments are fetched using Reddit's API. Utilizing a Language Model (LLM) with Retrieval Augmented Generation (RAG), the system evaluates comment relevance and filters out irrelevant comments. Sentiment analysis on filtered comments provides insights into public sentiment towards policies. The method is computationally efficient and integrates both retrieval-based and generation-based approaches. By aggregating insights, this approach informs policy refinement and generation based on public feedback.**

I. Problem Statement And Motivation:

We aim to develop an Information Retrieval System (IRS) web application, called 'Public Opinion Analysis for Government Policy Makers,' tailored for government officials responsible for public policy. This tool will gather public sentiment and feedback from Reddit regarding implemented policies. With the help of LLM using Retrieval Augmented Generation (RAG) techniques, it will analyze the Reddit comments, extracting public ideas and concerns aiding in policy refinement and development. Our project serves as a vital link between the public and policymakers, gathering and delivering the concerns and problems voiced by people on Reddit directly to government officials. By doing so, we bridge the gap between citizens and policymakers, facilitating the resolution of issues related to government policies.

II. Literature Review:

[2] The paper presents an automated approach for sentiment analysis and opinion mining, using data aggregation, natural language processing, and visualization techniques. It utilizes tools like the Stanford Parser and SentiWordNet to process and score user sentiments, creating graphical representations through the Google Chart API. Challenges include data quality and non-standardized expressions, while future work suggests self-learning mechanisms and dynamic data collection improvements. The system has potential applications for manufacturers, retailers, and policy makers.

[1] The research proposes a web scraping method using Facebook and Twitter Developers APIs, coupled with regular expressions, to tackle information overload, redundancy, and relevancy issues in social media. By extracting and filtering data based on user preferences, the method presents structured information in a database. The paper highlights the importance of web scraping in data extraction and references related tools. The implemented algorithm successfully addresses the challenges, offering potential for further development into a userfriendly application with automated scraping capabilities.

[4] The paper suggests a method for evaluating public opinion on government policies like Ganga rejuvenation and state separation via online forums, using sentiment analysis and natural language processing. It categorizes opinions by age and gender to aid policy formulation, employing machine learning and lexicon-based approaches. The system involves user registration, opinion collection, sentiment analysis, and graphical result visualization, enhancing democratic participation and policy making.

[5] The Paper adds to sentiment analysis literature by examining public sentiment towards Indian Government policies using Twitter data. It employs established methodologies for data extraction, pre-processing, and sentiment classification in R programming. By focusing on specific policies like Article 370 and the New Motor Vehicle Act 2019, the study provides insights into public perceptions of government initiatives. Its emphasis on graphical representation and comparative analysis aligns with trends in visualizing sentiment analysis results. Overall, the paper contributes to understanding public sentiment towards government policies and underscores the relevance of sentiment analysis in policy evaluation.

[3] The Paper presents the analysis of public opinion data related to the novel corona virus pneumonia epidemic. The study collected and analyzed comments from social network platforms and media sources to understand various views and attitudes towards the epidemic situation. Mathematical techniques such as natural language processing, sentiment analysis, topic modeling, network analysis, and statistical analyses were likely used to extract insights from the data. The research aims to provide a comprehensive understanding of public perceptions and interactions concerning the epidemic, highlighting the importance of analyzing online discourse during public health crises.

III. Novelty

Unlike existing approaches that mainly analyze sentiment, our project goes beyond by extracting actionable insights from public comments on Reddit. We not only provide sentiment statistics but also provide public ideas and the most frequently occurring concerns directly relevant to government policies offering our end-user i.e. policymakers valuable real-time feedback for informed decision-making.

IV. METHODOLOGY

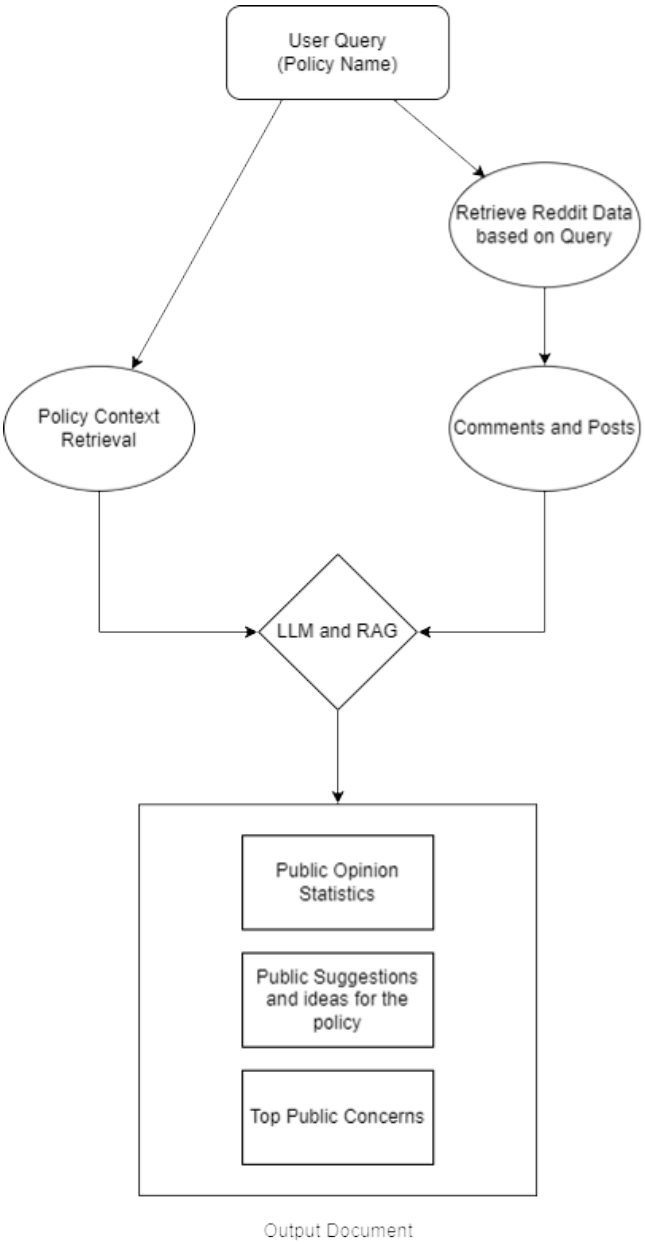

Fig. 1: Flow Chart

- User Input and Policy Selection: Users select a policy name from a dropdown menu on the web application interface. The selected policy name is used as a query to fetch relevant comments from Reddit. Implement a smart autocomplete feature to assist users in accurately selecting policy names. This feature suggests complete policy names as user's type, improving the accuracy of policy selection and reducing the risk of confusion.

- Data Collection from Reddit: Utilize Reddit's API to fetch public posts and comments related to the selected policy. Gather a sufficient number of comments to ensure comprehensive coverage of public opinions.

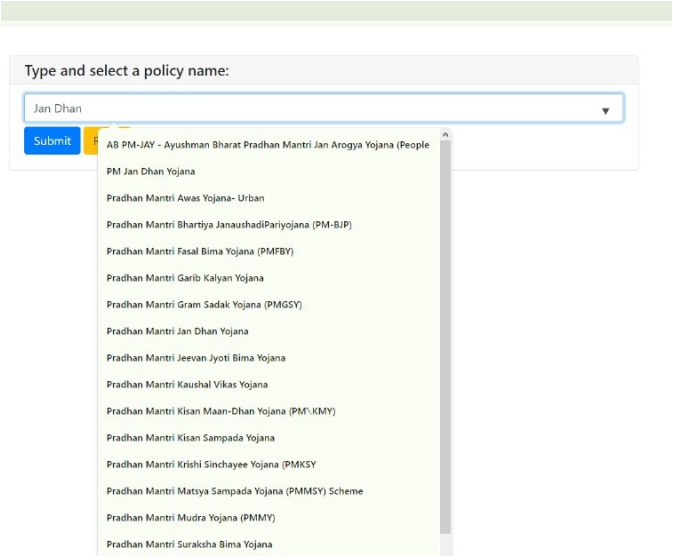

Fig. 2: Dropdown Menu For Policy Names

Fig. 3: Policy Input

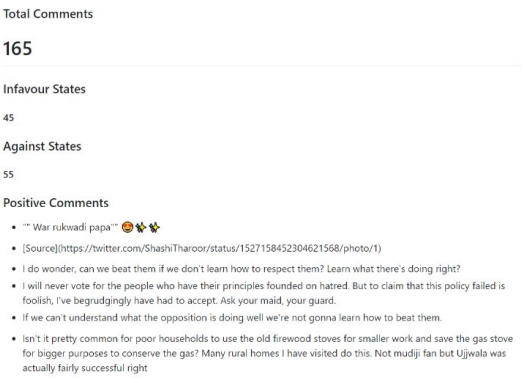

Fig. 4: Positive Comments

- Policy Context Retrieval: Retrieve context and details about the selected policy from trusted websites or sources. This information will provide background and context for understanding the policy being discussed in the Reddit comments.

- Language Model Evaluation with RAG: Utilize Gemini, a Language Model (LLM), enhanced with Retrieval Augmented Generation (RAG), to analyze the relevance of comments to the selected policy. RAG combines the capabilities of both retrieval-based and generation-based approaches, bringing an Information Retrieval (IR) component into the system.

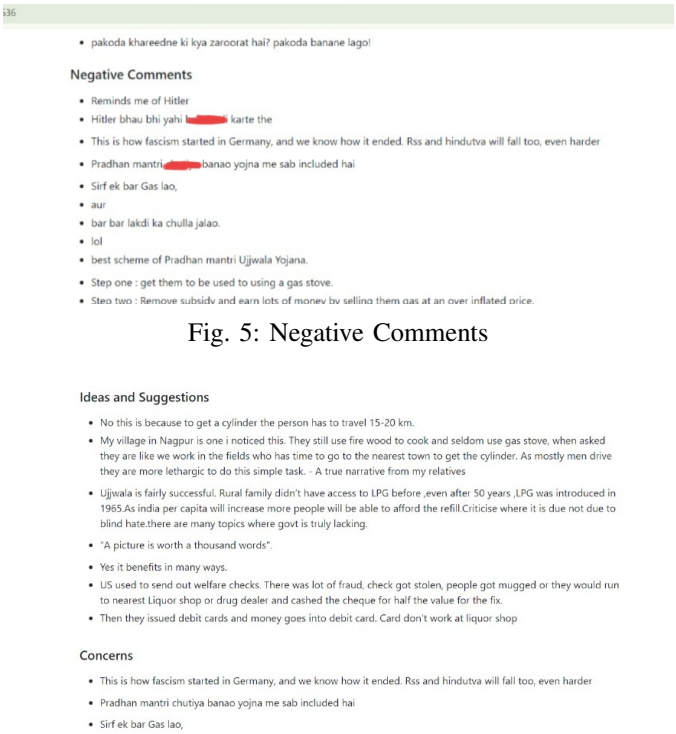

Fig. 5: Negative Comments

Fig. 6: Idea Suggestion and concern

This integration allows Gemini to leverage preexisting policy context retrieved from trusted websites while also generating responses, enhancing the understanding of comment relevance within the policy context. Evaluate each comment's semantic relevance to the policy context retrieved earlier using RAG. This process enhances efficiency by filtering out irrelevant comments and focusing on those that contribute meaningfully to the policy discourse. RAG offers several advantages over traditional fine-tuning approaches. Firstly, it is computationally inexpensive compared to fine-tuning the model on GPU using transformers. This efficiency is crucial for processing large-scale datasets commonly found in social media platforms like Reddit. Additionally, RAG's ability to integrate both retrieval and generation mechanisms enable a more comprehensive understanding of the comment's relevance within the policy context, leading to more accurate analyses and insights. Perform sentiment analysis on the filtered comments to determine public sentiment towards the policy.

- Policy Generation from Comments: Aggregate insights from the analyzed comments to generate new policies or refine existing ones. Identify recurring themes, concerns, or suggestions expressed in the comments. Use these insights to inform policymaking decisions and potentially adapt policies based on public feedback.

V. Database

We have collected policy names from many websites, preprocessed them and created a csv file named 'GovtPolicyList.csv'. We are using this csv file for the list of policy names.We are utilizing the Reddit's API to fetch public posts and comments related to the selected policy. Gather a sufficient number of comments to ensure comprehensive coverage of public opinions. We are retrieving the context and details about the selected policy from trusted websites and sources and storing it into json file named government policies.json'. This information will provide background knowledge and context for understanding the policy being discussed in the Reddit comments.

VI. Code

A. Previous:

https://github.com/gaurav-deswal/PolicyEchoSphere

B. Current:

https://github.com/Kirtanshah2303/IRProject

VII. Evaluation

By Manual Evaluation

```
Policy Name : Pradhan Mantri Jan Dhan Yojana
Precision : 0.35294117647058826
Recall : 0.75
F1_score : 0.48
Accuracy : 0.5357142857142857
Confusion Matrix
```

	Predicted Positive	Predicted Negative
Actual Positive	6	2
Actual Negative	11	9

Fig. 7: Pradhan Mantri Jan Dhan Yojana

```
Policy Name : Pradhan Mantri Kisan Samman Nidhi (PM Kisan)
Precision : 1.0
Recall : 0.625
F1_score : 0.7692307692307693
Accuracy : 0.7
Confusion Matrix
```

	Predicted Positive	Predicted Negative
Actual Positive	5	3
Actual Negative	0	2

Fig. 8: Pradhan Mantri Kisan Samman Nidhi Yojana

VIII. Contribution

All the team memebers have contributed equally. Gaurav and Kirtan worked on LLM model selection, model preperation and backend part. Jayshil and Amit worked on data preprocessing, created csv and json for further use and frontend/UI part. Ankur and Rohit worked on data gathering from different trusted sorces, project report and PPT

IX. Conclusion

Our research presents a robust framework for analyzing public sentiment towards policies by integrating smart autocomplete, Reddit's API, and Gemini with Retrieval Augmented Generation (RAG). Through the utilization of natural language processing and sentiment analysis techniques, we offer policymakers a comprehensive understanding of public perceptions, facilitating informed decision-making and adaptive policy development. This approach harnesses the power of technology to enhance democratic participation and responsiveness to societal needs.

References

[1] Lusiana Dewi, Meiliana, and Alvin Chandra. "Social Media Web Scraping Using Social Media Developers API and Regex". In: *Procedia Computer Science* 157 (2019), pp. 444–449. DOI: 10.1016/j.procs.2019.08.237.

[2] Pooja Kherwa et al. "An Approach towards Comprehensive Sentimental Data Analysis and Opinion Mining". In: *IEEE International Advance Computing Conference*. 2014.

[3] Xin Liu et al. "Public Opinion Analysis on Novel Coronavirus Pneumonia and Interaction With Event Evolution in Real World". In: *IEEE Transactions on Computational Social Systems* 8.4 (2021), pp. 1042–1051. DOI: 10.1109/TCSS.2021.3087346.

[4] Amit Pandhare and Abrar Alvi. "Public Opinion Mining For Effective Policy Making". In: *International Journal of Research in Advent Technology* 3.4 (2015). ISSN: 2321-9637.

[5] Pankaj Verma. "Mining Public Opinion on Indian Government Policies Using R". In: *International Journal of Innovative Technology and Exploring Engineering* 9 (2020). DOI: 10.35940/ijitee.C8150.019320.