2025 6th International Conference on Communications, Information, Electronic and Energy Systems (CIEES)
26 – 28 November, 2025, Ruse, Bulgaria

Benchmarking Machine Learning Techniques in Under-Resourced Contexts: Analysis of Public Perceptions of Government Policies from South Karnataka Reddit Discourse

Varnica Sharma
Dept. of Applied Statistics and Data Science, Prasanna School of Public Heath Manipal Academy of Higher Education
Manipal, Karnataka, India
varnicasharma1515@gmail.com

Aman Tripathi
Dept. of Applied Statistics and Data Science, Prasanna School of Public Heath Manipal Academy of Higher Education
Manipal, Karnataka, India
amantripathi3273@gmail.com

K.M. Kavitha
Dept. of Applied Statistics and Data Science, Prasanna School of Public Heath Manipal Academy of Higher Education
Manipal, Karnataka, India
karimbikavitha@gmail.com

***Abstract*—Understanding public perception is crucial for the successful implementation of government policies. This study employs established Natural Language Processing (NLP) techniques to gauge public sentiment of government policies in Indian context and identify key themes related to policies in the South Karnataka region of India. Traditional machine learning techniques such as Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Random Forest (RF) along with deep learning techniques such as DistilBERT were explored in uncovering public opinion on government policies under very low training data conditions. Latent themes in the public discourse were identified using K-Means Clustering and BERTopic. Our findings reveal a clear performance trade-off for the highly nuanced task of perception mining, as all the chosen models under-performed under low-data settings with SVM emerging as the best performing model with a 41.5% accuracy slightly higher than the deep learning model DistilBERT with an accuracy of 40.4%. For thematic clustering, K-Means produced highly coherent and separable topics with 94.6% accuracy, demonstrating its effectiveness in identifying keyword-driven themes, whereas BERTopic identified more semantically nuanced topics that were less distinct.**

Keywords—public perception mining, government policy analysis, low-resource NLP, reddit discourse analysis, sentiment analysis, machine learning benchmarking, DistilBERT, K-Means clustering, topic modeling

I. Introduction

Effective governance in the public sector hinges on the alignment between policy design and public acceptance. In the digital age, social media platforms have emerged as invaluable, albeit noisy, sources of unfiltered public opinion. This study harnesses the semi-anonymous, discussion-oriented platform Reddit to analyze citizen discourse surrounding Indian government policies in South Karnataka, India, a region characterized by rapid urbanization and evolving public service demands. We conduct a comparative study of two distinct Machine Learning (ML) paradigms: the traditional one which relies on statistical text features, and deep learning leveraging contextual embeddings for uncovering public perceptions on governmental policies. By evaluating the versatility of chosen models on the dual tasks of emotion classification and thematic clustering, we aim to provide a nuanced understanding of public sentiment and offer a critical perspective on the models employed for such analysis exclusively in under-resourced contexts of public policy perception under very low training data conditions.

The study presented in this paper might be viewed as a benchmarking application of established machine learning and deep learning techniques to a novel, under-resourced data context: public discourse on government policies originating from Reddit users in South Karnataka, India. Our work does not introduce novel machine learning models, but rather provides a crucial proof-of-concept for the effective deployment of current methodologies in analyzing a politically and geographically distinct digital community. Existing literature on political opinion mining in India has predominantly focused on large-scale national sentiment or metropolitan areas. There is significant gap in studies that isolate and analyze the opinions of regional, linguistically diverse communities such as those in South Karnataka. Reddit discourse provides a level of anonymity and depth often absent in other social media platforms, potentially offering unfiltered views on sensitive government policies. The primary research gap this study addresses is the absence of an established machine learning performance baseline for analyzing public discourse in this under-resourced, regional context of public policy perception. While global benchmarks exist, their utility in predicting sentiment within a unique, low-resource dataset of public policy perception remains untested. The novelty of this research lies in its empirical contribution—establishing the first comparative performance benchmark of machine learning models for topic modeling and sentiment analysis of public policy perception, a localized dataset of South Karnataka Reddit comments.

II. Motivation

A government policy's success is determined by the target citizens' perception, reception, and realized benefit, apart from the quality of its design and implementation. Indian public policy is characterized by a crucial disparity between official policy intent and the prevailing public discourse concerning its operational success and real-world impact. This study is an attempt to identify and analyze the divergence between the formal goals and desired outcomes of a policy (policy intent) and the way it is discussed, interpreted, and understood by the public (public discourse). In the said context, the contributions of the work presented in the paper include:

979-8-3315-0119-8/25/$31.00 ©2025 IEEE

- Showcase how a combination of established ML and NLP techniques (SVM, Term Frequency-Inverse Document Frequency (TF-IDF), etc.) can be used to derive meaningful insights from a unique social media dataset on Indian policies.
- Provide a comparative analysis of how existing techniques (SVM vs. DistilBERT) can be effectively applied to a specific, underexplored dataset on Indian policies, thus serving as a useful benchmark for future research in this area.
- Gain new knowledge and understanding about public discourse on Indian policies, which was previously unavailable.
- Furthermore, these findings offer a critical, low-resource performance baseline for policymakers and NLP practitioners working on regional Indian public opinion.

III. Review of NLP for Policy and Sentiment Analysis

Analysis of public sentiments across various domains has evolved significantly over the past two decades. Early sentiment analysis methods used handcrafted features to represent text. Models like SVM and NB relied on techniques such as n-grams and TF-IDF, but struggled to capture complex linguistic nuances like sarcasm and contextual meaning [1]. Specifically in the health domain, studies employing Twitter data to monitor public sentiment during the COVID 19 pandemic, often observed that public emotion shifted rapidly in response to government announcements and health outcomes [1].

The rise of deep learning, particularly transformer-based models like BERT, has revolutionized sentiment analysis by providing a richer understanding of semantic context [2]. Unlike earlier methods, these models could interpret complex arguments and subtle opinions that are not explicitly positive or negative [3]. Topic modeling techniques, such as the transformer-based BERTopic, have also emerged as powerful tools for identifying key themes in large corpora without supervision [4].

A major challenge in public policy is the disconnect between official goals and on-the-ground realities. For example, policies aiming for universal immunization may overlook regional infrastructure disparities, creating implementation barriers and resource mismatches. This gap is common in low and middle-income countries [World Health Organization (WHO) 2015], highlighting the need for methods that can capture grassroots sentiment. Similarly, analyses of education policies, such as India's National Education Policy (NEP) 2020, reveal gaps between goals and challenges (e.g., insufficient teacher training) [5]. For instance, an analysis of NEP 2020 Twitter data identified shifting public sentiment [6]. These parallels underscore the need for methodologies that help understand intent-implementation disparities across sectors [WHO 2015].

Existing literature demonstrates that NLP methodologies, validated in education and health policy contexts, provide a robust framework for bridging intent-implementation gaps. Synthesizing social media data offers policymakers a dynamic toolset for responsive, equity-focused reform [7]. Addressing cultural and demographic biases further enhance the model's scalability across diverse public health landscapes [WHO, 2015] [8]. A comparative study using machine learning for policy-related sentiment reinforces the value of benchmarking different models on a specific, topical dataset [9]. A highly relevant study employing BERT-based model to analyze public sentiment on a specific policy issue in the Indian context highlights the value of using modern NLP techniques on social media data to inform policymakers [10].

While a significant body of NLP research exists for analyzing political discourse in English and other major languages, there's a growing need to apply these techniques to under-resourced contexts. This includes analyzing social media discourse in specific regions, such as the focus on South Karnataka in this study, and dealing with challenges like code-mixing (English and regional languages) and limited data availability. There is a shift towards hybrid methods that combine the strengths of different techniques. For example, unsupervised topic models like BERTopic are used to identify key themes, which can then be fed into supervised models for detailed sentiment or classification tasks. This allows for both broad exploratory analysis and specific, targeted insights. Our study builds on the prior works [1, 4, Table I] by providing a direct comparative analysis in the under-resourced context of regional Indian policy discourse.

IV. Methodology

To understand the public perception on governmental policies we initially began by procuring the data set sourced from Reddit. To uncover public opinion on available policies, classification was performed followed by clustering for identifying latent themes. The workflow illustrating the steps from data collection through classification and identifying themes is summarized in (Fig. 1).

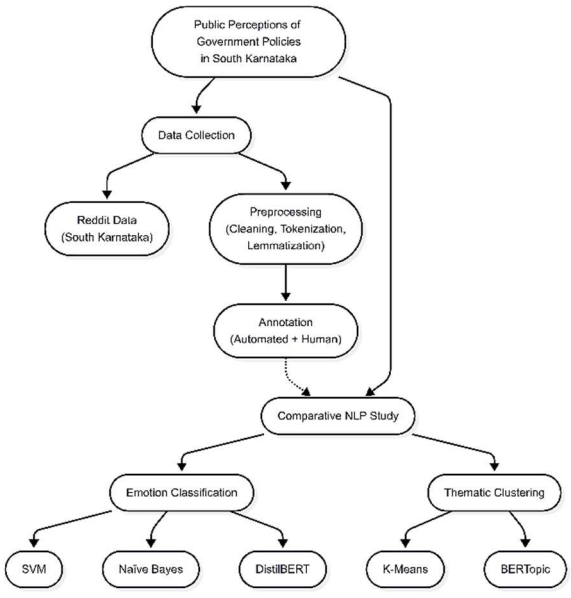

Fig. 1. Workflow illustrating data collection through analysis.

A. Text Acquisition and Corpus Building

To procure the dataset on public discourse on Indian policies, 735 Reddit comments were scraped using the PRAW API[1] from Reddit. The final CSV file consisted of five columns including "Comment Text", "Article Title", "Article URL", "Article Publication Date", and "Region". The region-specific responses were derived from subreddit, with region specific scraping centered on South Karnataka[2]. The

comments predominantly were in English with occasional local language terms and emojis. On average, the comment size was approximately 61 words, and the median size was 37 words. The overall vocabulary size was 6,700. Common terms included “Karnataka”, “government”, “people”, reflecting the regional and political focus. For example, the term “Karnataka” appeared 368 times and “government” was seen 276 times across the corpus. Emoji and Kannada characters were observed infrequently in a few comments. Approximately 461 comments were tagged to Bengaluru, 202 to “Other South Karnataka”, 66 to “Mysuru”, and the rest from nearby districts. These statistics indicate a moderately sized, topical dataset of social-media text concerning public health and policy.

For the unsupervised cluster analysis, the afore-said dataset was merged with a corpus of 461 comments with search specifically focusing on public health policies to form an expanded dataset of 1196 comments, increasing its diversity and robustness. The comments scraped from selected YouTube videos and Reddit posts majorly focused on public health policies. Over 500 comments were initially scraped and then filtered to remove spam, emojis, links, and irrelevant text using natural language pre-processing techniques. Analyzing the dataset we observe that the word “government” is the most frequently mentioned, indicating that users place primary responsibility for healthcare outcomes on the state.

Words like “people”, “citizen”, and “pay” reflect a grassroots concern: how individuals experience, pay for, or are affected by policies. The term “hospital” is notably high, showing concern around infrastructure and availability. “Private” appearing suggests comparisons or dissatisfaction with the public healthcare system, or a shift toward privatization. The mention of “insurance” implies active discussions on health financing and coverage mechanisms, especially around “Ayushman Bharat”. Words such as “government”, “govt”, “state”, “India” suggest high public scrutiny of government actions and decisions. Likewise, presence of words such as “hospital”, “scheme”, “card”, “doctor”, “treatment”, “private” point to concerns about the delivery of health services, access to care, and performance of public vs private sectors. Word Frequency plots and Word Clouds in (Fig. 2) and (Fig. 3) visually depicts the themes and contents within the text.

Fig. 2. Word cloud for social media comments focussed on public health policies.

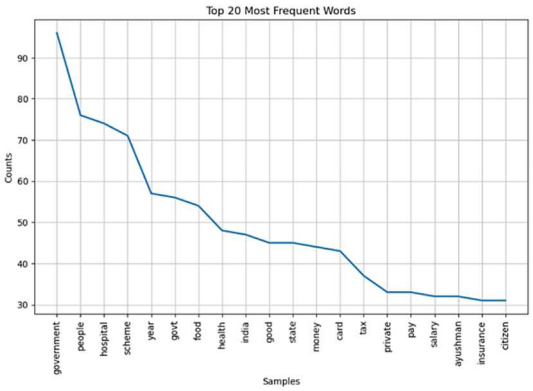

Fig. 3. Word frequency distribution plot for social media comments focussed on public health policies

B. *Text Pre-processing*

The procured comments were cleaned and prepared using a standard data pre-processing pipeline involving tokenization, lowercasing, stop word and punctuation removal, and lemmatization. Pre-processing was carried out using spaCy and NLTK toolkit. We additionally filtered URLs, punctuations and extraneous characters using regular expressions.

C. *Segregation of Public Perceptions*

We approached the task of analyzing public perceptions as a classification problem in combination with topic modeling and clustering for uncovering latent themes.

1) Lexicon-based Emotion Labeling: We first generated emotion labels for the 735 comments in the corpus using the lexicon-based NRCLex library[3]. This unsupervised step leverages a predefined dictionary of emotional terms. This enabled each comment a dominant emotion label “anger”, “joy”, “sadness”, “fear”, “surprise” and so forth. 80% of the labeled dataset was used as training set and 20% as test set.

2) Feature Extraction: To classify public responses, we extracted TF-IDF features using unigrams and bigrams from the cleaned comment text.

3) Classification: We trained and evaluated five traditional machine learning models such as SVM, DT, NB, KNN and RF. In parallel, we evaluated a pre-trained transformer model DistilBERT[4] [12] via HuggingFace’s pipeline to label each comment’s emotion.

Each classical model was trained on TF-IDF (1–2-grams; min_df = 5; max_df = 0.80) vectors using a stratified 80 / 20 train–test split (seed = 42) of the 735 NRCLex-labeled comments. Accuracy and weighted (wtd.) Precision, Recall, and F1-Score were computed on the held-out test set using scikit-learn (classification_report). The weighted average was chosen to provide a more representative performance measure due to the inherent class imbalance in the NRCLex-labeled emotion distribution.

Recognizing the limitations of pre-trained models on specialized data, we established a rigorous validation protocol. A human-annotated ground truth dataset was created by having three human raters independently label a random sample of 100 comments. Based on inter-annotator agreement “true” emotion for each comment was determined by a majority vote. The performance of the Large Language Model (LLM) baseline (DistilBERT) was then quantitatively

validated against this 100-comment human gold set (three-rater majority labels), followed by a qualitative error analysis to understand its specific strengths and weaknesses.

D. Thematic Clustering and Topic Modeling

In addition to the segregation of comments based on perceptions, we further considered identifying the latent themes underlying the perceptions to understand stakeholder opinions on governmental policies. On the merged dataset of 1196 comments, we compared two clustering techniques to identify latent topics in the discourse. We explored K-Means and BERTopic [13] models and are as detailed in the subsections that follow.

1) K-Means Clustering: We represented the text using TF-IDF vectors of uni- and bi-grams. To focus on thematically informative words, we filtered the vocabulary based on principles from Zipf's Law, setting a minimum document frequency of 5 and a maximum of 0.80.

2) BERTopic Modeling: As a modern alternative, we used BERTopic5, a transformer-based model that clusters comments based on the semantic similarity of their sentence embeddings, allowing for the discovery of more nuanced topics. Using BERTopic we extracted the top keywords of each topic to characterize main themes.

V. Results and Discussion

A. Emotion Classification Results

The task of classifying nuanced human emotions proved to be highly challenging for all chosen models. The unsupervised lexicon-based emotion distribution is visualized in (Fig. 4), which reveals an initial class imbalance. Table I summarizes the comparative performance of all models on the test set.

TABLE I. Performance of Classification Models.

Model	Precision (wtd.)	Recall (wtd.)	F1-Score (wtd.)
Traditional Models (on NRCLex Labels)			
SVM	0.30	0.41	0.34
DT	0.36	0.36	0.35
NB	0.28	0.35	0.25
KNN	0.31	0.23	0.16
RF	0.24	0.33	0.27
LLM (Validated on Human Labels)			
DistilBERT	0.46	0.40	0.37

Following prediction, the LLM's resulting distribution is shown in (Fig. 5) , contrasting sharply with the reference ground truth. The DistilBERT model, when evaluated against the human-annotated ground truth, achieved a validated accuracy of 40.4%. The pre-trained LLM did not significantly outperform the simpler traditional method suggesting that without domain-specific fine-tuning, even powerful general models struggle to interpret regional discourse.

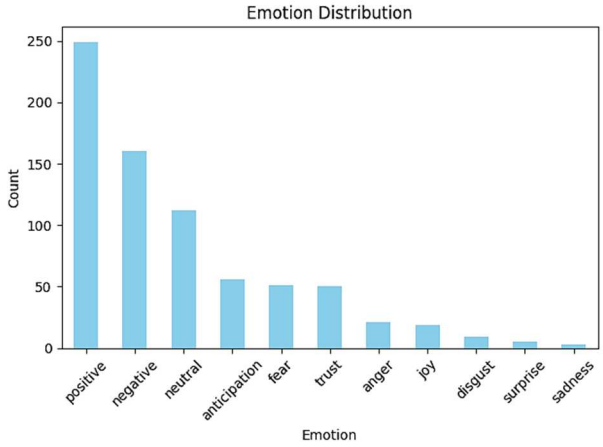

Fig. 4. Emotion Distribution from NRCLex Labels (Original Dataset).

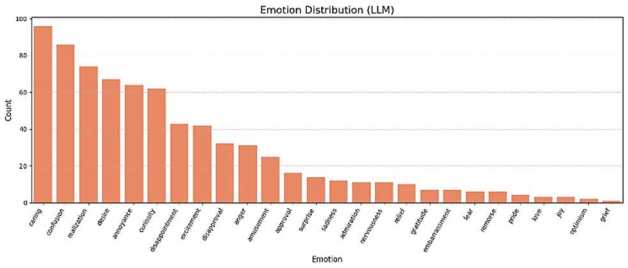

Fig. 5. Emotion Distribution from LLM Predictions (Original Dataset).

1) Error Analysis in Classification

A qualitative error analysis, detailed in Table I, provides insight into the LLM's behavior. The model correctly identified straightforward expressions of "confusion" and "gratitude". However, it faltered on comments where negative sentiment was conveyed indirectly, misinterpreting a comment expressing "disappointment" about water shortages as "confusion", and a comment expressing "anger" at government failure as "disappointment".

TABLE II. Error Analysis of LLM Predictions vs. Human Labels.

Evaluation (Comment Text)	Actual Emotion	Predicted Emotion
I'm not sure what to think anymore...	confusion	confusion
I'm so grateful for the people who are...	gratitude	gratitude
Karnataka has highest rainfall udupi yet...	disappointment	confusion
This government is gone bonkers...	anger	disappointment

B. Thematic Clustering Results

The optimal number of clusters was determined to be 5 using the Elbow Method, which identifies the point of diminishing returns in cluster inertia as depicted in (Fig. 6). The structure of the resulting clusters was visualized using Principal Component Analysis (PCA), and their coherence was evaluated by training an SVM to predict the generated cluster labels.

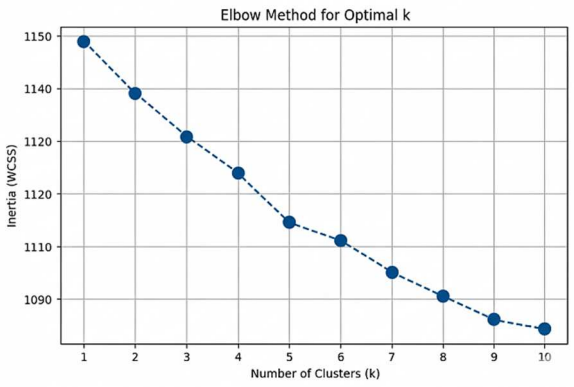

Fig. 6. Elbow plot justifying k=5 for K-Means.

Fig. 7 shows the distribution of documents across each of the 5 clusters identified using the K-Means clustering.

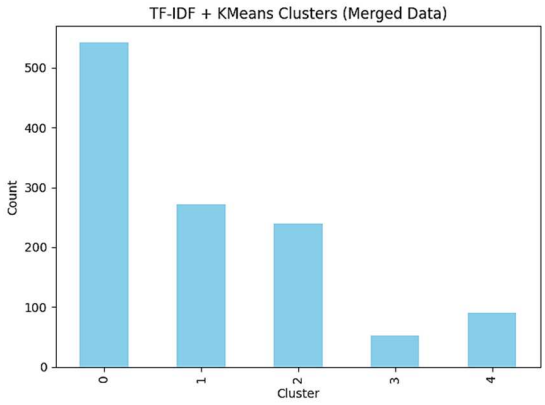

Fig. 7. Distribution of documents across the 5 K-Means clusters.

Table III summarizes the top 25 unigrams and selected bigrams identified for each of the 5 clusters using the K-Means Clustering. Table IV depicts the top terms identified using the BERTopic technique.

TABLE III. TOP TERMS FROM K-MEANS CLUSTERING.

Cluster	Top Terms
0	Unigrams: health, good, food, people, govt, scheme, like, india, hospitals, healthy, card, deleted, getting, money, great, time, don, protein, know, mandya, need, yes, scss, ayushman, salary. Bigrams: -
1	Unigrams: government, people, just, state, tax, central, like, money, don, india, work, companies, way, pay, private, need, make, good, point, bjp, years, city, karnataka. Bigrams: central government, state government
2	Unigrams: karnataka, kannada, language, state, people, just, north, like, hindi, south, bangalore, states, english, want, mysore, don, india, yes, learn, culture, bjp, land, hate, govt. Bigrams: north karnataka
3	Unigrams: mysuru, mysore, airport, city, don, posts, bengaluru, rich, big, years, real, post, new, young, bangalore, people, great, agree, golden, details, decade, like, develop, money, development. Bigrams: -
4	Unigrams: bengaluru, city, karnataka, urban, bangalore, https, namma, post, traffic, south, try, www, com, road, come, days, just, kannada, india, rural, did, places. Bigrams: namma bengaluru, karnataka bengaluru, https www

TABLE IV. TOP TERMS FROM BERTOPIC.

Cluster	Top Terms
0	karnataka, state, bjp, people, guy, media, tn, government, just, north, states, bihar, districts, trying, away, politicians, didnt, like, ka, news, good, auto, bangalore, development, india
1	hospitals, hospital, private, doctors, patient, card, ayushman, people, treatment, insurance, healthcare, patients, government, money, private hospitals, medical, health, blood, dont, ayushman card, know, covered, scheme, pay, india
2	food, protein, healthy, eat, indian, foods, sugar, diet, products, health, good, cooked, nutrition, india, eggs, indians, high, meat, stomach, fruits, rice, like, daily, cook, carbs
3	language, kannada, hindi, learn, karnataka, english, languages, south, speak, spoken, like, dont, kodava, indian, just, culture, south indian, people, tongue, tulu, professional, public, mother tongue, learning, mother
4	government, protests, just, protest, doing, lapse, fuck, work, action, question, instead, security lapse, fuck government, public, central, president, local government, pockets, people, politicians, central government, protesting, elected, power

While testing the coherence of the K-Means clusters by training an SVM to predict them, the model achieved an exceptionally high accuracy of 94.6%. This suggests that the clusters are linearly separable and thematically consistent. The PCA visualization in (Fig. 8) further illustrates these distinct groupings. In contrast, the SVM accuracy on BERTopic labels was much lower at 52.5%.

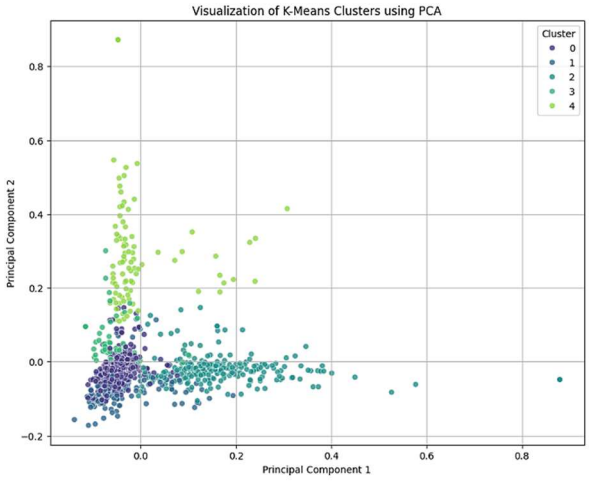

Fig. 8. Visualization of K-Means clusters using PCA.

C. *Limitations and Constraints*

Our work solely relies on analysis of public perceptions sourced from online discourses. This might underrepresent demographics with limited internet access or digital literacy, such as older adults, low-income individuals, and those in rural areas. Their perception of public policies, particularly those affecting healthcare, are crucial, nevertheless might go unheard. As data acquisition was sourced from South Karnataka and specific subreddits, it might not accurately reflect the opinions of the entire region or country. As the scraped comments were predominantly in English, with occasional local language terms, the analysis might not reflect the perceptions of majority of population communicating in regional languages. The use of Reddit and YouTube as data sources limits our analysis to comments from individuals who

actively use these platforms, which may not be a representative cross-section of society. The high frequency of terms, such as, “government” and “hospital” might reflect the concerns of an already engaged audience rather than the general population's primary issues. Thus, the dataset possibly overrepresents extreme or highly engaged viewpoints, while a silent majority with moderate or indifferent opinions is ignored, indicative of self-selection bias.

VI. Conclusion

This study conducted a comparative analysis of Machine Learning methods to dissect public discourse on health policies in South Karnataka, leveraging a unique dataset of approximately 1,200 comments from South Karnataka Reddit discourse. In an under-resourced, very low training data conditions, primary success of our work is measured by two key contributions: the rich thematic insights uncovered and the performance benchmark established for machine learning application in this regional setting.

Our validated results yielded a key insight: a powerful, pre-trained DistilBERT model (40.4% accuracy) performed comparably to a much simpler lexicon-based SVM (41.5% accuracy) for emotion classification, underscoring the critical importance of domain-specific validation. For thematic analysis, we found that K-Means produced highly coherent and separable clusters (94.6% SVM accuracy), indicating that keyword-based topics were more distinct than the semantically richer but overlapping topics from BERTopic.

By applying and comparing established machine learning and deep learning techniques such as SVM, DT, NB, KNN, RF, and DistilBERT we have established the first performance baseline for sentiment analysis and topic classification of public discourse on health policies within the South Karnataka Reddit corpus. While we acknowledge that our study is constrained by the small dataset size, this benchmark is essential for any subsequent, larger-scale studies in this domain, providing a necessary proof-of-concept for the feasibility of these methods in a low-resource context. While a pre-trained model like DistilBERT is effective, its performance can be significantly enhanced by fine-tuning it on the specific dataset. Fine-tuning a transformer model like BERT on this local, domain-specific data is a critical next step that could significantly improve emotion and topic classification accuracy, considering the feasibility of larger corpus of Indian public policy discourse.

While local language terms were occasionally observed in our corpus, with multilingual embeddings, employing multilingual models like SLM-RoBERTa, IndicBERT or Mbert [14], pre-trained on text from multiple languages including a wide range of Indian languages, would allow the model to more accurately process comments that mix English and Kannada or other regional languages rather than restricting to English parts. Cross-regional comparison would further enable us to generalize our findings to vast majority of the population in India, which primarily communicates in regional languages.

We envision a more robust system that is both interpretable and highly accurate by adopting a hybrid framework that merges the structured, keyword-based insights from traditional methods like TF-IDF with the deep contextual understanding of transformer models. Extending our analysis with multimodal data by including images or news links referenced by comments would further improve reliability and depth of public opinion insights, helping tailor communication strategies and policies more effectively. Temporal analysis capturing how sentiment and topics evolve over time with the new policy announcements is yet another potential area for exploration.

References

[1] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: Approaches, datasets, and future research," Applied Sciences, vol. 13, no. 7, p. 4550, 2023.

[2] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1253, 2018.

[3] G. Aadil and D. Samad, "Traditional or deep learning for sentiment analysis: A review," Multidiszciplináris tudományok, vol. 12, no. 1, pp. 3–12, 2022.

[4] M. Esh, "Sentiment analysis in chatgpt interactions: Unraveling emotional dynamics, model evaluation, and user engagement insights," Technical Services Quarterly, vol. 41, no. 2, pp. 160–174, 2024.

[5] K. C. Karthikeyan C, "Emerging challenges in implementation and practice of nep (national education policy) for teachers, students and stakeholders," Jan. 2021.

[6] J. Salam, "Draft national education policy (nep), 2019 and jingoistic nationalism," The People’s Chronicle, vol. VII, 2019.

[7] V. Anoop and S. Sreelakshmi, "Public discourse and sentiment during mpox outbreak: an analysis using natural language processing," Public Health, vol. 218, pp. 114–120, 2023.

[8] X. Gong and J. Wen, "The text analysis of national education policy 2020," Journal of Education, Teaching and Social Studies, vol. 5, p. p95, Feb. 2023.

[9] D. S. Zainulabdeen, M. Çevik, and M. M. Abdulrazzaq, "A comparative study of classification algorithms for sentiment analysis of covid 19 vaccine opinions using machine learning," in 2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2024, pp. 1–6.

[10] P. Khot, M. Vishwakarma, and V. Shukla, "Bert-based sentiment analysis of indian covid-19 tweets for policy making," Iconic Research And Engineering Journals, vol. 7, no. 8, pp. 26–33, 2024.

[11] O. Ütük Bayılmıs and S. Orhan, "Decoding digital labor: A topic modeling analysis of platform work experiences," MDPI, vol. 13, no. 9, p. 819, 2025.

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," ArXiv, vol. abs/1910.01108, 2019.

[13] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," arXiv preprint arXiv:2203.05794, 2022.

[14] R. Pathak and A. Kumar, "Aspect-based sentiment analysis in hindi language by ensembling pre-trained mbert models," Electronics, vol. 10, no. 21, p. 2641, 2021.