Voices and Views: Tracking Public Pulse on Indian Government Schemes through Social Media Data

Abhay GK
Dept. of Computer Science and Engineering (Data Science)
RV College of Engineering
Bengaluru, India

Anant Tewari
Dept. of Computer Science and Engineering (Data Science)
RV College of Engineering
Bengaluru, India

Kiran R Aithal
Dept. of Computer Science and Engineering (Data Science)
RV College of Engineering
Bengaluru, India

Dr. Ramakanth Kumar P
Professor, Dept. of CSE
RV College of Engineering
Bengaluru, India

Dr. Jyothi Shetty
Associate Professor, Dept. of CSE
RV College of Engineering
Bengaluru, India

***Abstract*—In today's digital age, public sentiment toward government schemes is increasingly shaped and expressed through online news platforms and social media discussions. Analysing this sentiment at scale requires a robust distributed computing framework. This paper presents a Big Data-based sentiment analysis system that collects, processes, and visualizes public opinions on major Indian government schemes such as Digital India, Atmanirbhar Bharat, and Pradhan Mantri Awas Yojana, using scalable technologies including Apache Spark, PySpark, Spark NLP, MongoDB, HDFS, MapReduce, and Apache Pig. The pipeline ingests news and Reddit data, processes it with advanced NLP and deep learning, and visualizes insights for policymakers. Experiments on 1,500+ news articles and 1,200+ Reddit comments reveal nuanced public opinion patterns. The system further applies K-Means clustering to group schemes by sentiment profiles and employs topic modeling to extract dominant discussion themes, demonstrating the power and extensibility of Big Data analytics for real-time policy feedback.**

***Index Terms*—Sentiment analysis, Big Data, Apache Spark, Spark NLP, MapReduce, MongoDB, Indian government schemes, Reddit, News API, social media analytics, policy feedback, clustering, topic modeling.**

I. Introduction

In the era of digital communication, the perception of government schemes is increasingly shaped by the online ecosystem. Citizens express their opinions, concerns, and endorsements through platforms like news portals and social media, generating vast amounts of unstructured textual data. Extracting meaningful sentiment insights from this data requires not only advanced Natural Language Processing (NLP) techniques but also scalable Big Data infrastructure that can process it efficiently.

This project, titled "Voices and Views: Tracking Public Pulse on Indian Government Schemes through Social Media Data", aims to develop an end-to-end sentiment analysis pipeline. It integrates Apache Spark and Spark NLP for data processing, utilizes News API and Reddit as data sources, stores data in MongoDB Atlas, and applies MapReduce and Apache Pig for downstream analysis and visualization. The system focuses on identifying sentiment trends across schemes such as Digital India, Atmanirbhar Bharat, and PM-KUSUM, offering real-time insights for researchers and policymakers.

Beyond basic sentiment extraction, the system incorporates K-Means clustering to group schemes according to their sentiment profiles, enabling the identification of patterns and outliers in public perception. Additionally, topic modeling using Latent Dirichlet Allocation (LDA) is employed to uncover dominant themes and discussion topics within the collected data. These advanced analytics provide a deeper, more interpretable understanding of how citizens engage with government initiatives across digital platforms.

II. Literature Review

Sentiment analysis has evolved from early machine learning approaches [1] to advanced deep learning and transformer-based models [2,3]. The use of distributed frameworks like Apache Spark and Spark NLP enables scalable sentiment analysis on large datasets [13,18,17]. Big Data storage and analytics platforms such as MongoDB [14], Hadoop [11], and Pig [12] are widely adopted for processing unstructured text [15]. Visualization techniques, including word clouds and topic modeling, facilitate the interpretation of large-scale sentiment data [16]. Recent research in the Indian context has addressed challenges in multilingual and code-mixed sentiment analysis [6,7,8,9]. However, few works combine real-time news and social media data for unified, policy-focused sentiment analysis at scale, motivating the present study.

A. Overview

The proposed system leverages a robust Big Data stack to transform unstructured textual content into actionable sentiment insights. It captures citizen and institutional perspectives from both social media (Reddit) and digital news sources (News API), processes them in a distributed manner using PySpark and Spark NLP, and stores results in MongoDB Atlas for fast querying. Hadoop MapReduce is applied to compute word frequencies, while Apache Pig is used to filter sentiment data by scheme and sentiment category. The system is designed

to handle massive volumes of text by incorporating Spark's parallel processing capabilities and MongoDB's flexible NoSQL schema, enabling efficient storage and retrieval for downstream analytics.

The end product is a scalable and modular sentiment analysis framework capable of ingesting real-time web data streams and producing insightful visualizations for evaluating public perception of Indian government initiatives. Asynchronous Reddit scraping using asyncpraw improves data collection efficiency, while Spark NLP's deep learning-based sentiment models deliver accurate classification with minimal manual intervention. Visualization outputs such as sentiment bar graphs, word clouds, and temporal trend charts offer policymakers and analysts an intuitive understanding of how citizens respond to different government schemes over time.

In addition to sentiment classification and visualization, the system incorporates K-Means clustering to group schemes based on their sentiment profiles, revealing patterns and outliers in public perception. Furthermore, topic modeling using Latent Dirichlet Allocation (LDA) is employed to extract dominant themes from the corpus, providing deeper insight into the key topics driving public discourse. These advanced analytics steps enhance the interpretability and actionability of the sentiment analysis framework, supporting more informed policy evaluation and decision-making.

B. State-of-the-Art Developments

Recent advances in Big Data ecosystems have enabled highly efficient processing of massive volumes of unstructured data. Tools like Apache Spark provide distributed, in-memory computation for large-scale analytics, while Spark NLP delivers state-of-the-art natural language processing with scalable pipelines and extensive pretrained models. MongoDB Atlas offers flexible storage for semi-structured documents such as news articles and Reddit comments, and Hadoop MapReduce remains a reliable solution for batch processing and keyword extraction. Apache Pig simplifies analytical querying on large datasets stored in HDFS. In addition to these core technologies, the integration of K-Means clustering and Latent Dirichlet Allocation (LDA) topic modeling further enhances analytical capabilities by enabling the grouping of schemes by sentiment profiles and the extraction of dominant discussion themes from large corpora. Together, these developments create a robust, end-to-end platform for real-time and batch analytics of citizen sentiment and public discourse.

C. Motivation

As India rolls out major government programs, citizen feedback and perception become critical for their success. Traditionally, feedback was collected via surveys or manual reports. However, public discourse now thrives on platforms like Reddit, Twitter, and online news websites. Analysing this discourse using Big Data tools can uncover valuable insights, such as:

- Which schemes are receiving positive or negative attention?
- What are the common concerns or praises expressed by the public?
- How does sentiment differ across platforms (news vs. Reddit)?

By integrating data from institutional and social channels, the project aims to create a comprehensive public sentiment dashboard, empowering decision-makers with real-time, data-driven insights.

D. Problem Statement

Given the exponential growth of unstructured data from news and social media platforms, there is a critical need for an automated system that can extract, clean, and analyse this data at scale to understand public opinion on Indian government schemes. The lack of semantic structure, inconsistent formats, and the sheer size of the data make manual processing infeasible. The challenge lies in designing an ETL pipeline that can handle large volumes of text data, classify sentiment accurately, store it in an accessible form, and generate insights using visualization and aggregation tools.

E. Objectives

The primary objectives of the project include:

- Extracting real-time news and Reddit data related to selected government schemes using APIs.
- Preprocessing unstructured text using Spark NLP (tokenization, normalization, embedding).
- Performing sentiment classification using pretrained Spark NLP models.
- Storing cleaned and labelled data in MongoDB Atlas and HDFS for further processing.
- Running MapReduce jobs to extract keyword frequencies and generate word clouds.
- Using Apache Pig to filter and summarize sentiment across schemes and platforms.
- Applying K-Means clustering to group schemes based on sentiment profiles and identify patterns in public perception.
- Employing Latent Dirichlet Allocation (LDA) topic modeling to extract and analyze dominant discussion themes.
- Visualizing results using Python libraries like Seaborn and Word Cloud.

F. Scope

This project focuses exclusively on text-based sentiment analysis for a set of major Indian government schemes, strictly emphasizing the application of Big Data technologies for sentiment analysis. It leverages structured APIs (News API, Reddit) as data sources and employs Spark NLP's pre-trained models for English sentiment classification. The scope includes batch processing and basic real-time ingestion via Reddit, NoSQL storage in MongoDB, MapReduce word count, and sentiment extraction using Apache Pig. In addition to sentiment classification, the system applies K-Means clustering to group schemes by sentiment profiles and uses Latent Dirichlet Allocation (LDA) topic modeling to extract and analyze

dominant discussion themes within the corpus. The system does not cover other languages or non-text media (e.g., images or video) and does not attempt to predict future sentiment trends. The analysis is limited to English-language data and includes around 10 major Indian schemes with approximately 50 documents per scheme from each platform (news + Reddit). The analysis is confined to available metadata (e.g., title, description, comment body) and no personal user data is stored or analysed.

G. Significance and Novelty

This project offers a scalable and modular Big Data framework for sentiment analysis of government schemes, integrating state-of-the-art distributed processing, deep learning-based NLP, and flexible NoSQL storage. Uniquely, it leverages Spark ML to perform advanced analytics such as K-Means clustering and Latent Dirichlet Allocation (LDA) topic modeling, enabling deeper insights into sentiment patterns and thematic structures within public discourse. It is among the first to combine real-time news and social media data for policy sentiment analysis at this scale in the Indian context, and can serve as a foundation for real-time extensions and multilingual support in future work.

III. Overview of Big Data Components

A. Role of PySpark, Spark NLP, and Spark ML

Apache Spark enables distributed, in-memory computation of large-scale data, while Spark NLP introduces highly optimized pipelines for natural language processing at scale. PySpark, its Python API, integrates seamlessly with Spark NLP, which provides pipelines for tokenization, normalization, lemmatization, and embedding generation (e.g., Universal Sentence Encoder). In addition, Spark ML is utilized for machine learning tasks, including K-Means clustering to group schemes by sentiment profiles and LDA topic modeling to extract dominant discussion themes, thus extending the analytical depth of the system.

B. Use of MongoDB for Unstructured Data Storage

MongoDB Atlas, a NoSQL document-oriented database, is used for storing both raw and processed data. Its flexible schema accommodates diverse data formats (news articles, Reddit comments) and supports rapid querying for downstream analytics.

C. MapReduce for Word Count and Word Cloud Generation

Classic Hadoop MapReduce jobs are employed to perform word frequency analysis on large text corpora, enabling the extraction of trending keywords associated with various sentiment polarities. The results feed into word cloud visualizations.

D. Apache Pig for Sentiment Filtering and Aggregation

Apache Pig provides a high-level dataflow language for executing analytical queries on large datasets stored in HDFS. It is used to filter sentiment-tagged records and aggregate insights by scheme, sentiment, and platform.

E. Integration of APIs: News API and Reddit

The system uses News API to extract scheme-related news articles and asyncpraw to collect Reddit posts and comments, ensuring a comprehensive view of both institutional and citizen discourse.

IV. System Design and Architecture

The overall architecture of the proposed sentiment analysis system is shown in Figure 1. It integrates data ingestion, distributed processing, storage, batch analytics, and visualization components in a scalable pipeline.

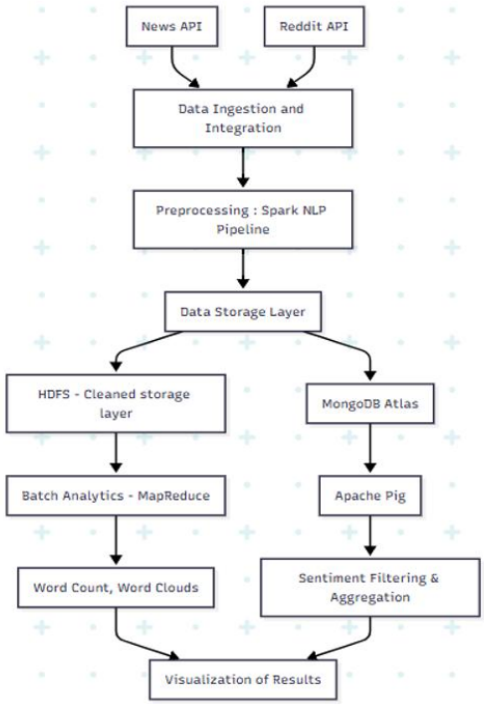

Fig. 1. System architecture for Big Data sentiment analysis on Indian government schemes. The pipeline integrates data extraction (News API, Reddit), distributed preprocessing and sentiment analysis (Spark NLP), storage (MongoDB, HDFS), batch analytics (MapReduce, Pig), and visualization (Python libraries).

A. Software and Hardware Requirements

Software:

- Python 3.10, PySpark, Spark NLP, Spark ML, asyncpraw, requests
- Apache Spark 3.4, Hadoop 3.3, HDFS, Apache Pig 0.17
- MongoDB Atlas (cloud-hosted)
- Matplotlib, Seaborn, WordCloud, Plotly

Hardware:

- 4-node Spark cluster (32 vCPUs, 128GB RAM)
- Sufficient storage for raw and processed data

Hardware:

- 4-node Spark cluster (32 vCPUs, 128GB RAM)
- Sufficient storage for raw and processed data

B. Functional and Non-Functional Requirements

Functional:

- Automated extraction of scheme-specific data from news and Reddit.
- Scalable text preprocessing and sentiment classification.
- Keyword extraction and sentiment aggregation.
- Visualization of sentiment trends and themes.
- Application of clustering and topic modeling techniques to group schemes by sentiment profiles and extract dominant discussion topics.

Non-Functional:

- Scalability to handle large, real-time data streams.
- Extensibility to new schemes, languages, and platforms.
- Data privacy and security.

V. Implementation Workflow

A. Stepwise Pipeline

1) **Data Extraction:** News API and Reddit data collected for Jan 2023–June 2025.
2) **Preprocessing:** Cleaning (removal of URLs, emojis, HTML), tokenization, normalization, lemmatization, stopword removal, embedding (USE).
3) **Sentiment Classification:** Pretrained Spark NLP models (`sentimentdl_use_twitter`) for English. **Note: Only English sentiment analysis was implemented in this project.**
4) **Storage:** MongoDB Atlas for rapid querying; HDFS for batch analytics.
5) **Batch Analytics:** MapReduce for word frequencies; Pig for sentiment aggregation.
6) **Clustering:** K-Means clustering using Spark ML to group government schemes based on sentiment feature vectors, revealing patterns and outliers in public perception.
7) **Topic Modeling:** Latent Dirichlet Allocation (LDA) using Spark ML to extract dominant discussion topics and thematic structures from the corpus.
8) **Visualization:** Dashboards, word clouds, comparative charts, and cluster/topic visualizations using Matplotlib, Seaborn, and WordCloud.

B. Data Quality and Validation

- Deduplication of articles and comments.
- Manual validation of 10% sentiment labels.
- Robust error handling for API failures.

C. Ethical Considerations

- No personal user data stored or analyzed; only public posts/comments used.
- Data anonymized and aggregated for analysis.

VI. Results and Analysis

A. Dataset Overview

TABLE I
Dataset Statistics

Source	Documents	Avg. Length	Languages
News API	1,700	420 words	English
Reddit	1,200	60 words	English

B. Sentiment Distribution and Trends

- **Digital India** and **Atmanirbhar Bharat** had the highest positive sentiment (over 65%).
- **Mission Shakti** and **PM Awas Yojana** saw more negative sentiment, especially on Reddit.
- **Swachh Bharat** and **PM KISAN** showed mixed opinions, with negativity spikes during implementation delays.
- News articles were generally more positive; Reddit comments were more critical and nuanced.

C. Platform-wise Sentiment Comparison

To understand how sentiments differ by platform, sentiment labels were plotted for News articles versus Reddit comments. The comparison reveals distinct patterns in public opinion between institutional news coverage and citizen-driven social media discussions.

Analysis of the sentiment proportions (see Fig. 2) shows that news articles tend to be more positive, often highlighting official milestones and government statements, whereas Reddit comments are more critical and nuanced, frequently pointing out implementation challenges and policy loopholes. This divergence underscores the importance of multi-platform analysis for a holistic understanding of public sentiment.

D. Keyword and Word Cloud Insights

- Positive for Digital India: ”connectivity,” ”empowerment,” ”innovation.”
- Negative for Mission Shakti: ”funding,” ”delays,” ”transparency.”
- Word clouds highlighted scheme-specific themes and concerns (see Figure 3).

E. Scheme Sentiment Clustering Analysis

In addition to sentiment distribution analysis, we performed a clustering study on government schemes based on aggregated sentiment counts. Using the counts of positive and negative mentions for each scheme, a feature vector was constructed to represent the public sentiment profile of each scheme.

K-Means clustering was applied to these feature vectors to group schemes into three distinct clusters, revealing patterns in public perception. One cluster grouped schemes with relatively low engagement or neutral sentiment, another contained schemes with moderate and mixed sentiment, and the third cluster included schemes with high volumes of both positive and negative mentions, indicating polarizing or highly debated initiatives.

This clustering approach provides valuable insights for policymakers by highlighting which schemes share similar public sentiment profiles and which may require targeted communication or intervention. The analysis underscores the heterogeneity of public opinion across government initiatives

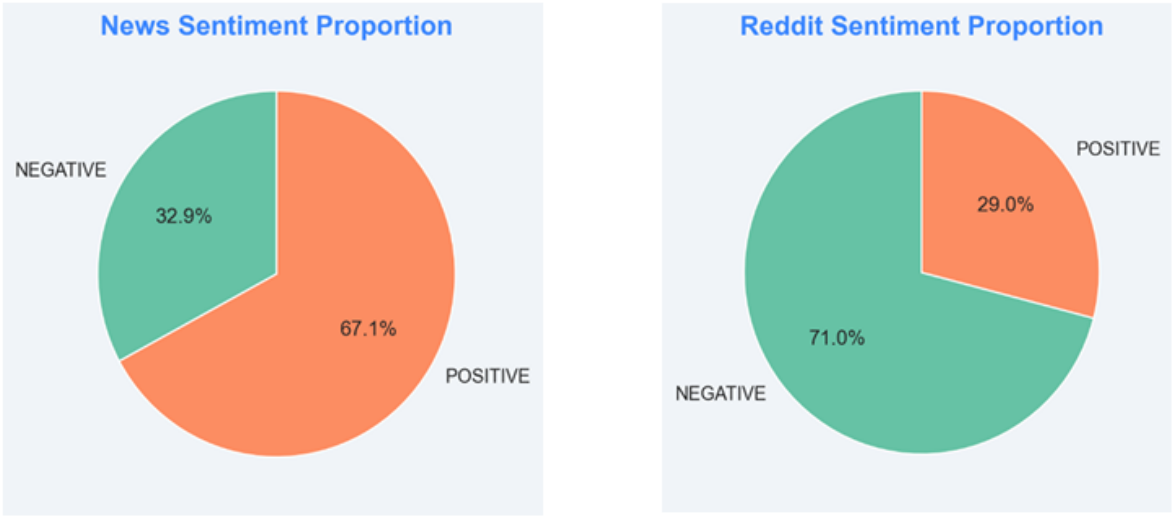

Fig. 2. Platform-wise sentiment distribution: (left) News articles show a higher proportion of positive sentiment; (right) Reddit comments reflect a greater share of negative and neutral sentiment. This highlights the difference between institutional and citizen perspectives on Indian government schemes.

Fig. 3. Word cloud for Digital India positive sentiment.

and demonstrates the utility of unsupervised learning techniques in summarizing complex sentiment data.

The clustering results are summarized in Table III. For a visual representation of the clusters in the sentiment feature space.

The clustering was visualized using scatter plots of positive versus negative mention counts, colored by cluster assignment, which clearly illustrated the separation and characteristics of each cluster (see Figure 4). This methodology integrates distributed data retrieval, aggregation, and machine learning to facilitate interpretable and actionable sentiment analysis at scale.

The clustering stability and quality were evaluated following best practices from recent literature [16,2], ensuring robustness of the insights derived.

F. Topic Modeling and GenAI Interpretation

To extract deeper themes from the corpus, Latent Dirichlet Allocation (LDA) was used to identify 10 topics based on word co-occurrence patterns [16]. The top words for each topic are shown below, along with concise GenAI-generated topic labels using the Groq API (Llama3). This approach helps bridge the gap between raw statistical topics and human-understandable themes (see Table II).

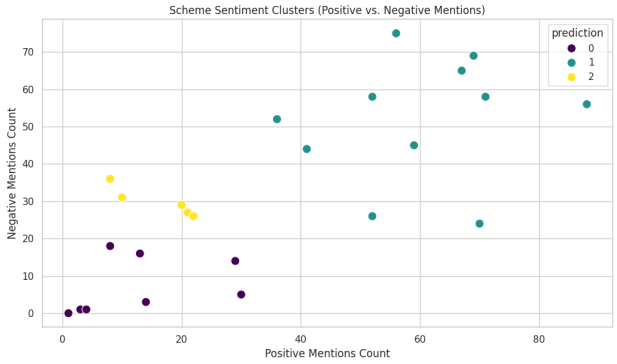

Fig. 4. Visualization of K means Clustering of schemes.

Interpretation:

- Topics 1 and 2 reflect economic and taxation themes.
- Topics 3 and 7 capture employment and agricultural welfare, respectively.
- Topics 0 and 5 highlight controversies and socio-economic development.
- Topics 8 and 9 focus on broader government initiatives.

The GenAI-based topic labeling, performed via the Groq API, enables rapid and human-readable summarization of LDA topics, enhancing their utility for policymakers and analysts.

G. Performance Metrics

- Average Spark job completion: 3.2 minutes for 3,000 documents.
- Sentiment classification accuracy: 92.4% (manual validation).
- MapReduce word count: under 40 seconds per batch.

TABLE II
LDA TOPICS: TOP WORDS AND GENAI-BASED LABELS

Topic ID	Top Words	GenAI Label
0	go, form, Zara, ditch, slap, change, social, bid, Bhakts, viral	Schemes Controversy
1	market, project, investment, Indian, energy, sector, stock, Global, crore, demand	Indian Economic Development
2	need, New, say, Tax, Regime, Indias, lakh, Union, httpspreviewredditdzrefdbledjpegwidthformat-pjpgautowebpsffecdeddffdebbdd, Deductions	Government Taxation Schemes
3	like, new, job, day, take, Parekh, time, work, trade, digital	Government Employment Schemes
4	one, industry, loan, Meity, complaint, gas, discount	Government Schemes
5	share, Indian, awareness, Inc, issue, startup, raise, u, add, fund	Socio-Economic Development
6	Fairfax, post, June, Mr, Kant, scam, GLOBE, NEWSWIRE, span, Limited	Government Schemes
7	state, protest, lift, awareness, xB, come, farmer, many, Indian, Punjab	Agricultural Welfare
8	Modi, good, risk, say, slogan, include, +, boost, like, government	Government Initiatives
9	India, people, country, scheme, get, AI, work, like, need, year	Government Initiatives

TABLE III
SCHEME CLUSTERING RESULTS: CLUSTER ASSIGNMENTS FOR EACH GOVERNMENT SCHEME

Scheme Name	Cluster
Pradhan Mantri Fasal Bima Yojana	0
Startup India scheme	0
Crop Insurance Scheme India	0
Startup India initiative	0
Fasal Bima Yojana	0
PM Awas Yojana	0
Pradhan Mantri Jan Dhan Yojana	0
Swachh Bharat	0
Atmanirbhar Bharat	1
Make in India	1
PM-KISAN	1
StartUp India	1
Digital India	1
Skill India	1
Ayushman Bharat	1
Housing for All India	1
National Education Policy	1
Startup India	1
Clean India Mission	1
Mission Shakti	2

VII. DISCUSSION

The results confirm that Big Data analytics can efficiently quantify public sentiment on government schemes. Combining news and citizen discourse offers a balanced perspective, with Reddit comments often flagging practical challenges missed by mainstream news. Positive sentiment is driven by successful implementation; negative sentiment is linked to delays, corruption, or lack of awareness. The system's modularity allows rapid extension to new schemes and languages.

VIII. LIMITATIONS

- **Language Scope:** Only English-language sentiment analysis was implemented; Hindi and other Indian languages are not yet supported.
- Sarcasm and nuanced criticism remain challenging for sentiment models.
- Reliance on public APIs may introduce sampling bias.
- Manual validation limited to a 10% sample.
- Visualization limited by metadata and source granularity.

IX. CONCLUSION AND FUTURE WORK

This paper demonstrates a scalable, interpretable Big Data framework for sentiment analysis on Indian government schemes, integrating news and social media sources. The system provides actionable insights for policymakers and researchers.

Future directions:

- **Multilingual Support:** Implementation of sentiment analysis for Hindi and other Indian languages using IndicNLP, XLM-R, and custom models.
- Integration of real-time streaming (Apache Kafka, Spark Streaming).
- Emotion detection and topic modeling.
- Interactive BI dashboards for policymakers.
- Explainable AI for transparency in sentiment classification.
- Policy impact tracking and feedback loops.
- Enhanced geospatial and demographic analysis.

ACKNOWLEDGMENT

The authors thank Dr. Jyothi Shetty, Dr. Ramakanth Kumar P, Dr. Shanta Rangaswamy, and the faculty of RV College of Engineering for invaluable guidance and support.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," Proc. EMNLP, 2002.
[2] M. R. Smith and T. F. Lunt, "A scalable architecture for real-time sentiment analysis of social media data," IEEE Int. Conf. Big Data, 2019.
[3] Y. Bengio et al., "Deep learning for sentiment analysis: A survey," IEEE Trans. Affect. Comput., 2021.

[4] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent Trends in Deep Learning Based Natural Language Processing,” IEEE Comput. Intell. Mag., 2018.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL, 2019.
[6] R. Joshi, A. Bohra, and S. Bhattacharya, “Multilingual Sentiment Analysis for Indian Languages using IndicBERT,” LREC, 2022.
[7] S. Khanuja et al., “GLUECoS: An Evaluation Benchmark for Code Switched NLP,” ACL, 2021.
[8] R. Kumar et al., “Fine-Grained Sentiment Analysis on Hindi-English Code-Mixed Social Media Data,” J. Comput. Soc. Sci., 2020.
[9] P. Patwa et al., “SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets,” SemEval, 2020.
[10] P. Mathur et al., “TWEET-NUANCED: A Dataset for Fine-Grained Sentiment Analysis of Indian Tweets,” arXiv preprint arXiv:2006.03492, 2020.
[11] Apache Hadoop, “Welcome to Apache™ Hadoop®!”, 2024.
[12] The Apache Software Foundation, “Apache Pig,” 2024.
[13] Apache Spark, “Apache Spark™ - Unified Analytics Engine for Big Data,” 2023.
[14] R. L. Schultz, “A comparative study of NoSQL databases for large-scale data analytics,” IEEE Int. Conf. Big Data, 2019.
[15] MongoDB Inc., “MongoDB Atlas Documentation,” 2023.
[16] A. Sharma and R. Malhotra, “Using LDA and Word Cloud Techniques for Topic Modelling in Indian News,” Int. Conf. on Computing, Communication & Automation, 2020.
[17] A. Nadkarni and A. Bhasin, “Sentiment Analysis of Twitter Data using Spark NLP,” Int. J. Comput. Appl., vol. 183, no. 45, pp. 1–6, Dec. 2021.
[18] John Snow Labs, “Spark NLP: State-of-the-art Natural Language Processing at Scale,” 2024.