MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 13 Issue 02 (2026) | Pages: 432-438

A Result Paper On CrowdPulse: From Trends to Insights

Kamini R. Mohite[1], Prasad S. Sangale[2], Aditya V. Bholane[3], Chaitanya M. Kalbhor[4], Shahid A. Shaikh[5]

[1,2,3,4,5]Department of Artificial Intelligence and Data Science, S. B. Patil College of Engineering, Indapur, Pune, India

[1]kaminimohite5151@gmail.com, [2]sangaleprasad2005@gmail.com, [3]bholaneaditya430@gmail.com, [4]chaintanyakalbhor42@gmail.com

[5]shahid.shaikh23@gmail.com

Peer Review Information

***Type**: Article*
***Received**: 22 March 2026*
***Revised**: 06 April 2026*
***Accepted**: 24 May 2026*
***Published**: 05 June 2026*

Abstract

The rapid advancement of Natural Language Pro- cessing (NLP) and transformer-based deep learning has created new opportunities to analyze and compare information nar- ratives across different media platforms. This paper presents CrowdPulse, a cross-platform intelligence system designed to compare news media coverage with public conversations on Reddit. The system collects real-time data from both sources and applies advanced NLP techniques including RoBERTa- based sentiment classification, achieving a validation accuracy of 94.02%, FASTopic topic modeling with an optimal coher- ence score of 0.485 at 15 topics, and Sentence-Transformer- based semantic similarity computation. The system further integrates an AI-powered daily digest using the Gemini 2.0 Flash model, narrative drift alerts, subreddit-level sentiment breakdown, and narrative framing tag generation. Results are presented through an interactive React-based web dashboard. The importance of understanding differences between media framing and public opinion is a major takeaway of this project.

Keywords: Sentiment Analysis; Topic Modeling; Transformer Models; Semantic Similarity; Social Media Analytics; Cross-Platform Intelligence; Narrative Analysis; RoBERTa; FASTopic; Natural Language Processing.

How to Cite This Article

Mohite, K. R., Sangale, P. S., Bholane, A. V., Kalbhor, C. M., & Shaikh, S. A. (2026). A result paper on CrowdPulse: From trends to insights. *Multidisciplinary Journal of Research in Engineering and Technology*, 13(2), 432–438.

Introduction

CrowdPulse is a cross-platform intelligent system that compares and analyzes structured news media content with unstructured public opinions on Reddit. In today's digital world, information spreads through two distinct channels: traditional journalism and online communities such as Reddit. Understanding how these two platforms construct different narratives about the same event is of significant importance for researchers, journalists, and policymakers alike.

CrowdPulse collects real-time data from Reddit using the PRAW API and from news portals using the NewsData.io API. It processes textual data and converts it into mean- ingful analytical results using various Artificial Intelligence tools including transformer-based sentiment analysis, topic modeling, semantic similarity computation, and named entity recognition (NER). The system unifies these tools into a single platform, enabling structured comparative analysis between media narratives and public opinion.

In recent years, the integration of deep learning models with NLP has enabled development of systems capable of detecting subtle differences in framing, tone, and emphasis across large-scale text corpora. Many studies have examined Reddit or news data in isolation; however, few systems provide an integrated pipeline for cross-platform comparison with real- time data ingestion. The CrowdPulse system addresses this gap by automating data collection, analysis, and visualization within a single cohesive framework.

The proposed platform performs automated Reddit and news article scraping, RoBERTa-based sentiment classification with CSV-level caching, FASTopic topic modeling, semantic sim- ilarity computation using cosine similarity, entity frequency analysis using SpaCy NER, and cross-platform narrative com- parison. Results are presented through an interactive React- based web dashboard featuring AI-generated daily digests, narrative drift alerts, subreddit breakdowns, topic velocity tracking from historical snapshots, cross-platform quote ex- traction, and framing tag generation using the Gemini large language model.

Literature Survey

Several researchers have explored sentiment analysis, topic modeling, and social media analytics in recent years. The following survey summarizes the most relevant works that inform the design of CrowdPulse.

Dash et al. [1] applied a BERT-based transformer for sen- timent classification of Reddit posts in peer-to-peer networks, addressing challenges posed by informal and slang-heavy content. The work demonstrated improvements in accuracy over traditional approaches but was limited by domain-specific dataset sizes. Future scope included larger datasets and sar- casm detection.

Babariya et al. [2] combined VADER sentiment scoring with LDA topic modeling on Reddit data, highlighting the limitations of single-model pipelines. Their work empha- sizes the benefit of multi-model approaches and motivates the integrated pipeline adopted in CrowdPulse. Future scope suggested hybrid transformer-topic models and real-time in- teractive dashboards.

Kedzierska et al. [3] used LDA to extract themes from large Reddit datasets, noting the difficulty of topic coherence at scale. The adoption of BERTopic and dynamic topic modeling was suggested as future work, directly motivating the use of FASTopic in the present system.

Guerra and Karakus [5] applied lexicon-based emotion scoring with topic modeling to Reddit posts during the Russia- Ukraine conflict, demonstrating the value of cross-domain sentiment analysis in geopolitical contexts. Extension to other socio-political events and real-time dashboards was identified as future work.

Kang et al. [6] employed Word2Vec embeddings and transformer-based topic modeling to study linguistic differ- ences between neurodiversity communities on Reddit, sup- porting the application of NLP techniques in social science research.

Zayats and Ostendorf [7] modeled Reddit discussions as hierarchical threaded conversations using a Graph-Structured LSTM, improving contextual understanding beyond flat text analysis. CrowdPulse leverages similar insights to enhance topic coherence across comment chains.

Xu et al. [4] analyzed public attitudes toward ChatGPT on Reddit using sentiment and topic analysis, demonstrating the value of cross-platform comparison for understanding public perception of emerging technologies.

Nwaoha et al. [8] tracked sentiment changes over time on Reddit using temporal LDA with sentiment trend analysis. This work directly informs the timeline and daily snapshot features implemented in CrowdPulse.

Limitations of Existing Work

Current systems for social media and news analysis exhibit several shortcomings that the CrowdPulse system aims to address.

Most existing platforms analyze Reddit or news data in isolation without providing a direct cross-platform comparison mechanism. Systems that do compare sources often rely on static or pre-collected datasets, lacking real-time data ingestion capabilities that are essential

for timely narrative analysis.

Sentiment analysis tools frequently rely on lexicon-based approaches such as VADER, which struggle with complex, contextually dependent expressions including sarcasm, irony, and the domain-specific informal language common on Reddit. Deep learning approaches that address these limitations are rarely integrated with topic modeling and entity recognition within a single end-to-end pipeline.

Topic modeling using standard LDA produces inconsistent results with short social media texts and does not capture semantic relationships between words. Approaches based on transformer embeddings offer improved coherence but are computationally expensive without optimization. Furthermore, existing systems generally lack features for narrative fram- ing analysis, drift detection between platforms, subreddit- level sentiment breakdown, and AI-generated interpretive sum- maries, all of which are implemented in CrowdPulse.

Motivation

In today's information-intensive environment, the same event is frequently reported through multiple lenses: traditional news media present institutionally framed narratives, while Reddit communities reflect crowd-driven, emotionally engaged public discourse. These two perspectives often diverge signif- icantly in sentiment, emphasis, and framing.

Understanding this divergence is critical for several stake- holders. Journalists can identify stories receiving dispropor- tionate attention on social media compared to formal news coverage. Researchers can study how media framing shapes or differs from public perception over time. Policymakers can assess public sentiment toward specific events or policy de- cisions. Businesses can monitor brand narratives and product sentiment across platforms in real time.

Manually analyzing hundreds of news articles and thousands of Reddit posts to identify these patterns is impractical. There is a strong need for an automated, intelligent system that can perform sentiment detection, topic discovery, semantic alignment, and narrative comparison at scale. This need gave rise to the concept of CrowdPulse, which leverages the power of Artificial Intelligence and Natural Language Processing to provide reliable, scalable, and user-friendly cross-platform media intelligence.

Proposed System

Problem Statement

Public discourse is represented differently on social media and news media. There is currently no comprehensive auto- mated system for comparing emotions, discourse patterns, and narrative framing between these two media types at scale. A proper analytical system is needed to quantify and visualize the gap between public perception and institutional media framing, and to make these insights accessible through an interactive interface.

System Architecture

The CrowdPulse architecture follows a layered design con- sisting of four primary layers as illustrated in Figure 1: the Data Layer, Processing Layer, Analytics Layer, and Presenta- tion Layer.

- The Data Layer interfaces with the Reddit API via PRAW and the NewsData.io News API to retrieve real-time posts and articles respectively. Collected data is stored as structured CSV files in the raw data directory, versioned with timestamps to support historical comparisons.
- The Processing Layer contains the Reddit scraper and news scraper modules. The Reddit scraper fetches hot posts from configured subreddits with a configurable post limit, extracting title, body text, score, comment count, upvote ratio, creation timestamp, and URL. The news scraper queries the NewsData.io API using topic-specific search terms, extracting title, description, content, publication date, source name, image URL, and article URL.
- The Analytics Layer performs all NLP and machine learn- ing computations, including text preprocessing, RoBERTa- based sentiment inference, FASTopic topic modeling, SpaCy named entity recognition, and Sentence-Transformer semantic similarity computation. This layer also includes the sentiment worker, which labels scraped CSVs once per scrape cycle and stores results back to disk, eliminating redundant inference on subsequent dashboard loads.
- The Presentation Layer consists of the React-based fron- tend dashboard served via Vite, communicating with the FastAPI backend through a RESTful API. The frontend ren- ders interactive charts, topic cards, sentiment visualizations, the AI digest, and the insights module using Recharts and Tailwind CSS.

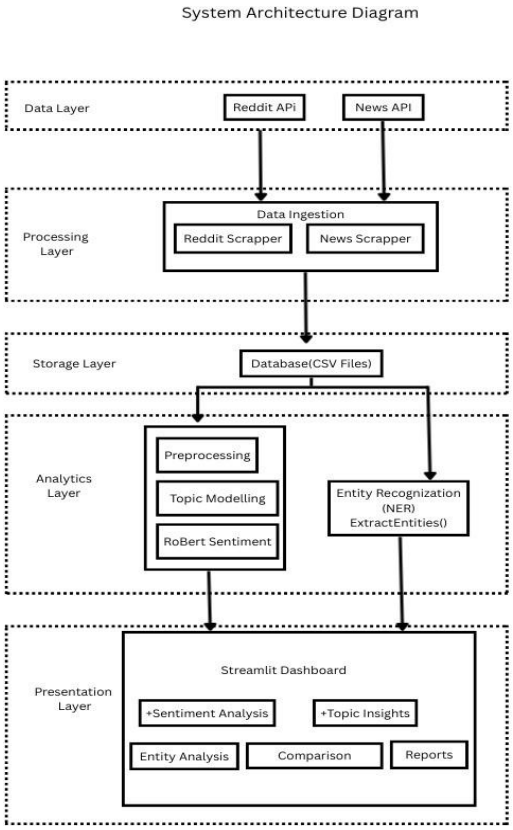

Fig. 1. *System Architecture Diagram*

System Workflow

he system workflow begins by checking data freshness, triggering automated Reddit and news scrapers only when necessary to ensure optimal performance. Collected text is preprocessed and analyzed using the RoBERTa sentiment classifier, FASTopic modeling, and spaCy NER to extract themes and entities. Semantic similarity is computed via Sentence-Transformers to quantify narrative divergence, while the Gemini 2.0 Flash model synthesizes high-level daily di- gests. Finally, all insights, including historical snapshots, are cached and served as structured JSON responses for interactive frontend visualization.

Key Analytical Features

The CrowdPulse system implements several intelligent fea- tures to assist users in identifying narrative patterns across different platforms:

- Narrative Drift Detection: For each discovered topic, the system computes the absolute difference between average news and Reddit sentiment scores. Topics with a drift score exceeding 0.3 trigger automated alerts on the dashboard to highlight major cross-platform divergences.
- Subreddit Sentiment Breakdown: The system groups Reddit data by specific subreddits for each topic. This reveals community-specific stances, contrasting institu- tional reports with niche perspectives found in specialized subreddits.
- AI-Powered Framing Tags: Utilizing the Gemini 2.0 Flash model, the system generates structured framing labels such as “Economic Anxiety” or “Government Distrust.” This provides a qualitative understanding of how each platform anchors its narrative.
- Topic Velocity Tracking: By analyzing seven-day his- torical JSON snapshots, the system tracks keyword fre- quency to label topics as Rising, Stable, or Fading, allowing users to monitor the evolution of media trends.
- Cross-Platform Quote Extraction: The system identi- fies and surfaces representative quotes from both news articles and Reddit threads based on high-confidence sentiment predictions, providing immediate context for the detected narrative.
- Source Reliability Metadata: Each analysis is supported by metadata including source name, publication date, and author credentials, ensuring transparency and enabling users to verify the origin of each narrative framing.

Implementation Details

Technology Stack

The complete technology stack used in the CrowdPulse system is summarized as follows. The backend is imple- mented in Python 3.x

using the FastAPI framework with Uvicorn as the ASGI server. Data manipulation is han- dled by Pandas and NumPy. The sentiment model uses HuggingFace Transformers with a locally stored fine-tuned RoBERTa checkpoint. Topic modeling uses the FASTopic library with the Topmost preprocessing pipeline. Semantic similarity uses the Sentence-Transformers library with the all-MiniLM-L6-v2 model. Named entity recognition uses SpaCy with the en_core_web_sm model. Vector search uses FAISS. User authentication uses JWT tokens with python- jose and bcrypt via passlib. The database layer uses SQLite with SQLAlchemy ORM. AI generation uses the Google Generative AI (Gemini) Python SDK.

The frontend is implemented in React 18 with Vite as the build tool, Tailwind CSS for styling, Recharts for data visualization, Axios for API communication, React Router v6 for navigation, Lucide React for icons, and Framer Motion for animations.

Results and Discussion

Sentiment Classification Performance

The RoBERTa-based sentiment classifier was evaluated after four training epochs on a labeled three-class sentiment dataset. Table I presents the full evaluation results.

Table 1. RoBERTa Sentiment Classification Performance

Metric	Value
Evaluation Loss	0.2230
Accuracy	94.02%
Weighted F1-Score	94.03%
Macro F1-Score	93.41%
Weighted Precision	94.04%
Weighted Recall	94.02%

Training loss decreased consistently across all epochs, con- firming effective model optimization. Validation loss stabilized at epoch 3 and slightly increased at epoch 4, identifying epoch

3 as the optimal checkpoint. Peak inference throughput of approximately 26 samples per second was achieved at this configuration. The RoBERTaForSequenceClassification archi- tecture with 12 transformer layers, 768 hidden dimensions, 12 attention heads, GELU activation, and 0.1 dropout demon- strated strong contextual understanding and high generaliza- tion capability across diverse news and Reddit text styles.

Topic Modeling Evaluation

Topic modeling quality was evaluated using the coherence score C_v. The optimal coherence of 0.485 was achieved with 15 topics. Lower topic counts (6–8) produced under- specified clusters, while counts above 16 introduced topic fragmentation. The FASTopic model demonstrated stable and interpretable keyword extraction across diverse corpora com- bining formal news language and informal Reddit discourse.

Semantic Similarity and Drift Analysis

Cosine similarity between news and Reddit topic embed- dings was computed as:

$$\text{drift} = |\bar{s}_{\text{news}} - \bar{s}_{\text{reddit}}| \qquad (2)$$

where $\bar{s}$ is the average sentiment score mapping POSITIVE to +1, NEGATIVE to −1, and NEUTRAL to 0. Topics with drift exceeding 0.3 were flagged as alerts. In testing on real scraped data, geopolitical topics consistently produced the highest drift scores, confirming that institutional news and public Reddit discourse frame sensitive events differently in both tone and emphasis.

System Interface Results

The complete system was successfully implemented and tested across all functional modules. The dashboard correctly renders KPI metrics, the AI daily digest, side-by-side news and Reddit article cards with thumbnail images, discovered topics with velocity badges,

similarity progress bars, sentiment distribution pie charts, and entity cloud visualizations.

The Insights module renders per-topic sentiment split bar charts, drift alert visualizations with horizontal bar charts, cross-platform quote cards with source links, topic velocity sparklines, subreddit sentiment stacked bars with top post titles, and Gemini-powered framing panels displaying tags, dominant frame descriptions, tone labels, and key actors.

The Search module supports on-demand query-based fetch- ing from both platforms with full sentiment annotation, entity extraction, and topic modeling applied to results. The Book- marks module correctly matches bookmarked topics against current data using sentence embedding similarity. The Time- line module renders historical daily snapshots with expandable topic, entity, and headline details.

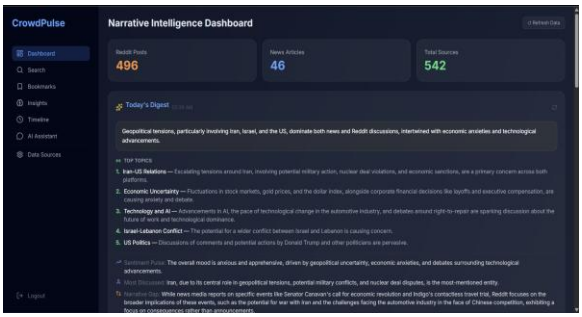

Fig. 2. *Home Dashboard View*

$$\text{Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \qquad (1)$$

where **A** and **B** are the Sentence-Transformer embeddings of news and Reddit topic keyword strings respectively. Narra- tive drift was computed as:

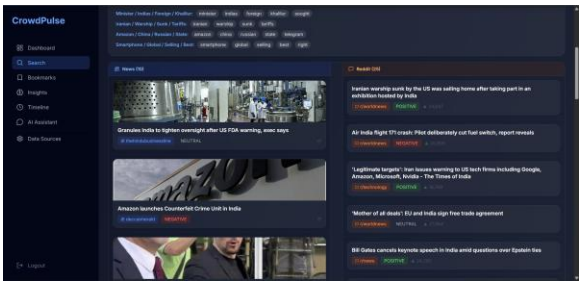

Fig. 3. *Topic Search Interface*

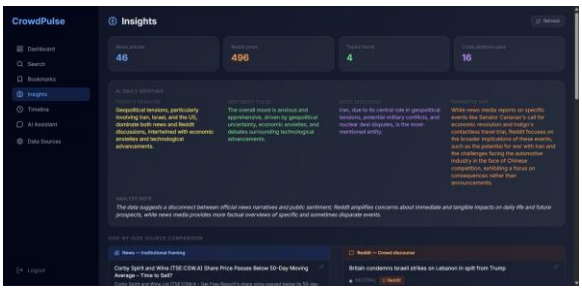

Fig. 4. *Comparative Analytics Module*

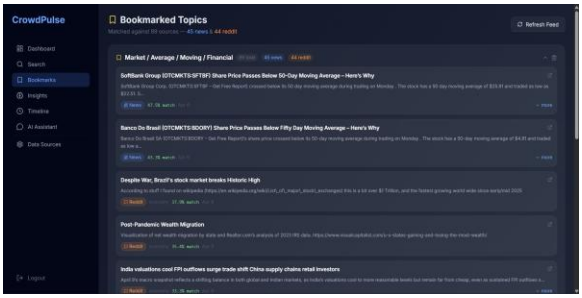

Fig. 5. *Personalized Feed and Bookmarks*

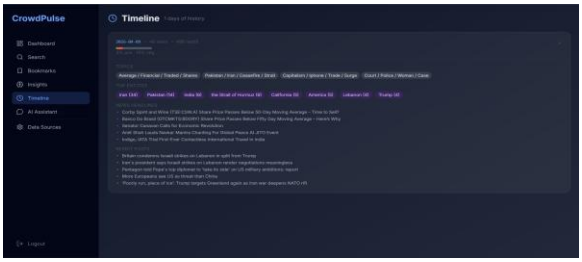

Fig. 6. *Historical Sentiment Timeline*

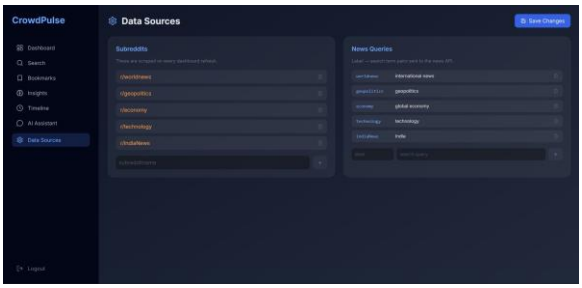

Fig. 7. *Source Reliability and Metadata*

Conclusion

CrowdPulse was developed as a cross-platform intelligence system to analyze and compare structured news media cover- age with public discussions on Reddit. The system success- fully integrates advanced NLP techniques including RoBERTa- based sentiment classification, FASTopic topic modeling, se- mantic similarity computation, named entity recognition, and AI-driven narrative generation within an interactive web-based dashboard built on a modern React and FastAPI stack.

The experimental results demonstrate that the RoBERTa sentiment classifier achieved a validation accuracy of 94.02% with strong macro and weighted F1-scores. Topic modeling optimization identified 15 topics as the most coherent con- figuration with a C_v score of 0.485. Semantic similarity and drift analysis quantitatively revealed alignment and divergence between media and public discourse, with geopolitical topics consistently showing the largest narrative gaps.

The system effectively highlights differences in tone, em- phasis, and thematic framing between institutional reporting and community-driven discussions through its narrative drift alerts, subreddit sentiment breakdowns, cross-platform quote extraction, topic velocity tracking, and Gemini-powered fram- ing tag analysis.

Overall, the project validates the feasibility of combining transformer-based deep learning models with real-time data pipelines to build scalable cross-platform intelligence systems. Future enhancements include multilingual sentiment and topic modeling support, real-time streaming data integration, cloud- based deployment, advanced misinformation and bias detec- tion, and graph-based narrative tracking across longitudinal datasets.

References

1. P. K. Dash et al., “Sentiment Analysis of Reddit Posts Using the BERT Model in Peer-to-Peer Networks,” in *Proc. Int. Conf. Intelligent Systems and Embedded Design (ISED)*, 2024.
2. D. Babariya et al., “Sentiment Analysis and Topic Modeling of Reddit Data,” *IEEE Access*, 2025.
3. M. Kedzierska et al., “Topic Modeling Applied to Reddit Posts,” *Lecture Notes in Computer Science*, Springer, 2023.
4. X. Xu et al., “Public Attitudes Towards ChatGPT on Reddit,” *arXiv*, 2024.
5. P. Guerra and A. Karakus, “Measuring Hope and Fear in Reddit Posts During Russo-Ukrainian Conflict,” *ICWSM*, 2023.
6. H. Kang et al., “Linguistic & Topic Analysis of Trends in ADHD vs Autism Reddit Communities,” *IEEE Access*, 2025.
7. V. Zayats and M. Ostendorf, “Conversation Modeling on Reddit Using a Graph-Structured LSTM,” *ACL*, 2017.
8. E. Nwaoha et al., “Longitudinal Sentiment Topic Modelling of Reddit Posts,” *J. Medical Internet Research*, 2024.
9. Y. Cai et al., “Public Sentiment about ChatGPT in Mental Health Discussions,” *arXiv*, 2024.
10. J. Li et al., “Sentiment Analysis and Topic Modeling on Reddit,” *IEEE Access*, 2025.
11. HuggingFace, “Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX,” 2024. [Online]. Available: https:// huggingface.co/transformers
12. Scikit-learn Developers, “Scikit-learn: Machine Learning in Python,” 2024. [Online]. Available: https://scikit-learn.org/