Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278- 5140
Volume 14 Issue 01, 2025

CrowdPulse: From Trends to Insights in Reddit Discussions

[1]Prof. Vinay S. Nalawade, [2]Aditya V. Bholane, [3]Chaitanya M. Kalbhor, [4]Prasad S. Sangale, [5]Shahid A. Shaikh
[1 2 3 4 5]Department of Artificial Intelligence and Data Science
S. B. Patil College of Engineering, Indapur, Pune, India
Email: [1]vinaynalawade2007@gmail.com, [2]bholaneaditya430@gmail.com, [3]chaitanyakalbhor42@gmail.com, [4]sangaleprasad2005@gmail.com, [5] shahid.shaikh23@gmail.com

Peer Review Information

Submission: 11 Sept 2025

Revision: 10 Oct 2025

Acceptance: 22 Oct 2025

Keywords

Reddit Analysis; Natural Language Processing (NLP); Sentiment Analysis; Data Visualization; Social Media Analytics; Trend Detection; Topic Modeling

Abstract

Reddit is a major source of user-generated content filled with discussions, opinions, and emotions. However, analyzing this data is challenging due to slang, sarcasm, and informal text. This study introduces **CrowdPulse**, a modular Reddit analysis system that processes and visualizes discussions into meaningful insights. The system includes modules for data collection, text cleaning, sentiment analysis, and visualization, offering transparency, scalability, and interactivity for researchers and businesses.

Introduction

Reddit generates vast amounts of unstructured discussions filled with slang, sarcasm, and noise, making analysis difficult. Existing tools provide only superficial insights and are often designed for other platforms such as Twitter.

To address these challenges, this paper introduces **CrowdPulse**, which integrates data collection, text cleaning, sentiment analysis, and visualization into a single analytical framework. The system is designed to transform unstructured Reddit discussions into actionable insights.

Contributions of this work:

1. Automated extraction and preprocessing of Reddit discussions.
2. Integration of sentiment and topic modeling techniques.
3. Visualization of patterns and insights using an interactive dashboard.
4. Scalable and modular system design for future expansion.

Problem Statement

Reddit discussions contain a wealth of user opinions and sentiments but present technical challenges due to informal language, slang, sarcasm, and rapidly changing topic contexts. The massive volume and velocity of posts make manual analysis impractical, while existing platforms typically lack support for precise trend detection and fine-grained emotion extraction in Reddit-specific formats. Recent real- world events (e.g., public health, political discourse, social community issues) generate thousands of relevant discussions, but patterns and insights remain hidden unless scalable, automated text analytics are applied. The motivation for this project is to enable robust, modular extraction of sentiment and topic trends from Reddit posts, supporting research, decision-making, and

platform transparency.

Literature Survey

The following table summarizes key research papers related to Reddit and social media analytics.

No.	Paper Title	Author(s) & Year	Problem Solved	Technique Used	Future Scope
1	Sentiment Analysis of Reddit Posts Using the BERT Model in Peer-to-Peer Networks	Dash et al., 2024	Capturing nuanced sentiment in peer-to-peer platform posts	BERT embeddings + fine-tuned classification	Transfer learning to Reddit sentiment contexts and domain adaptation
2	Performance Evaluation of Reddit Comments Using Machine Learning and NLP Meth- ods in Sentiment Analysis	Zhang et al., 2024	Compare traditional classifiers (Naive Bayes, SVM) with transformer-based models (BERT, RoBERTa, GPT) for sentiment classification	Naive Bayes, SVM, transformer models	Use transformer models for richer contextual analysis
3	Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora	Hamilton et al., 2016	Improving accuracy of sentiment analysis using domain-specific lexicons	Unsupervised induction of sentiment lexicons	Extend to more domains
4	Conversation Modeling on Reddit Using a Graph-Structured LSTM	Zayats and Ostendorf, 2017	Capturing conversational structure for Reddit discussions	Graph-LSTM sequence modeling	Dialogue systems and argument mining
5	GoEmotions: A Dataset of Fine-Grained Emotions	Demszky et al., 2020	Building and classifying a dataset for fine-grained emotion detection	Dataset creation, fine-grained emotion classifiers	Expand fine-grained emotion datasets
6	Topic Modeling Applied to Reddit Posts	Kędzierska et al., 2023	Uncovering hidden topics in Reddit posts	LDA topic modeling	Domain adaptation, new platforms
7	Sentiment Analysis and Topic Modeling Study of Food Security Discussions on Social Media	Molenaar et al., 2024	Analyzing food security discourse on Reddit	Sentiment and topic modeling	Study cross-platform food discourse
8	Measuring Hope and Fear in Reddit Posts During Russo-Ukrainian Conflict	Guerra and Karakuş, 2023	Quantifying public emotion (hope/fear) during conflict	Lexicon scoring + LDA	Socio-political events, emotional polarity
9	Linguistic & Topic Analysis of Trends in ADHD vs Autism Reddit Communities	Kang et al., 2025	Tracking semantic convergence of discussions about ADHD and autism	Word2Vec + BERT-based topic modeling	Tracking evolving neurodiversity dialogues
10	Sentiment Analysis and Topic Modeling of Reddit Data	Babariya et al., 2025	Insights on returning-to-office sentiment data	VADER & LDA on Reddit comments	Scaling to broader workplace domains

Proposed System

CrowdPulse implements a modular pipeline for Reddit data analytics:

- **Automated Data Collection**: Reddit posts/comments are gathered using the PRAW API, with support for filtering by subreddit, keywords, or time frame.
- **Preprocessing and Cleaning**: Raw text is

cleaned to remove HTML tags, special characters, and stopwords; tokenization and lemmatization provide standardized inputs for downstream mod- eling.

- **Sentiment and Topic Modeling**: Multiple sentiment analysis techniques (VADER, trans- former models) and topic discovery methods (LDA, BERTopic) are orchestrated to enable dual insights (emotion and trend).
- **Interactive Dashboard**: A dashboard visualizes extracted sentiment scores, trending topics, user engagement, and temporal patterns, using libraries such as Plotly and Dash.

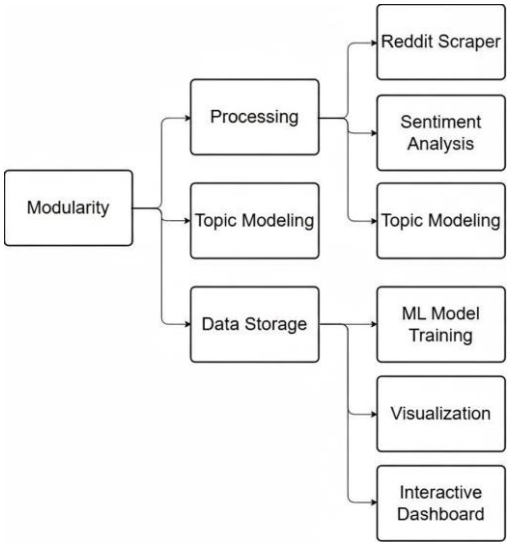

Figure 1: Proposed System Diagram

Methodology

The CrowdPulse methodology integrates several steps for effective Reddit analytics:

- **Text Preprocessing**: Each Reddit post undergoes regex-based cleaning, normalization, tok- enization, stopword removal, and lemmatization to improve downstream model accuracy.
- **Model Training and Analysis**: Sentiment analysis leverages both lexicon-based (VADER) and deep learning transformer models fine-tuned for social media text. Topic modeling uses LDA and BERTopic for multi-granular trend mapping.
- **Dashboard and Visualization**: Insights from NLP models are visualized on an interactive dashboard (Plotly/seaborn), supporting temporal comparison, cross-topic monitoring, and gran- ular emotion analysis across Reddit communities.
- **Evaluation and Feedback Loop**: The pipeline supports iterative improvement and retraining, to adapt models to new Reddit language or event trends.

Conclusion

CrowdPulse offers a robust analytics pipeline for Reddit discussions using modern NLP and visual- ization techniques. The system enables scalable sentiment analysis, topic modeling, and interactive dashboarding on complex social data. Automated preprocessing and deep models help uncover valu- able insights from unstructured posts. This framework supports improved research, decision-making, and social media transparency.

References

Dash, P. K., et al. "Sentiment Analysis of Reddit Posts Using the BERT Model in Peer-to-Peer Networks." Proc. Int. Conf. on Intelligent Systems and Embedded Design (ISED), 2024.

Zhang, X., et al. "Performance Evaluation of Reddit Comments Using Machine Learning and NLP Methods in Sentiment Analysis." arXiv preprint, arXiv:2405.16810, 2024.

Hamilton, W. L., et al. "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora." EMNLP, 2016.

Zayats, V., and Ostendorf, M. "Conversation Modeling on Reddit Using a Graph-Structured LSTM." ACL, 2017.

Demszky, D., et al. "GoEmotions: A Dataset of Fine-Grained Emotions." ACL, 2020.

Kędzierska, M., et al. "Topic Modeling Applied to Reddit Posts." LNCS, Springer, 2023.

Molenaar, A., et al. "Sentiment Analysis and Topic Modeling Study of Food Security Discussions on Social Media." J. Med. Internet Res., 26(1), 2024.

Guerra, P., and Karakuş, A. "Measuring Hope and Fear in Reddit Posts During Russo-Ukrainian Conflict." ICWSM, 2023.

Kang, H., et al. "Linguistic & Topic Analysis of Trends in ADHD vs Autism Reddit Communi- ties." IEEE Access, 2025.

Babariya, D., et al. "Sentiment Analysis and Topic Modeling of Reddit Data." IEEE Access, 2025.