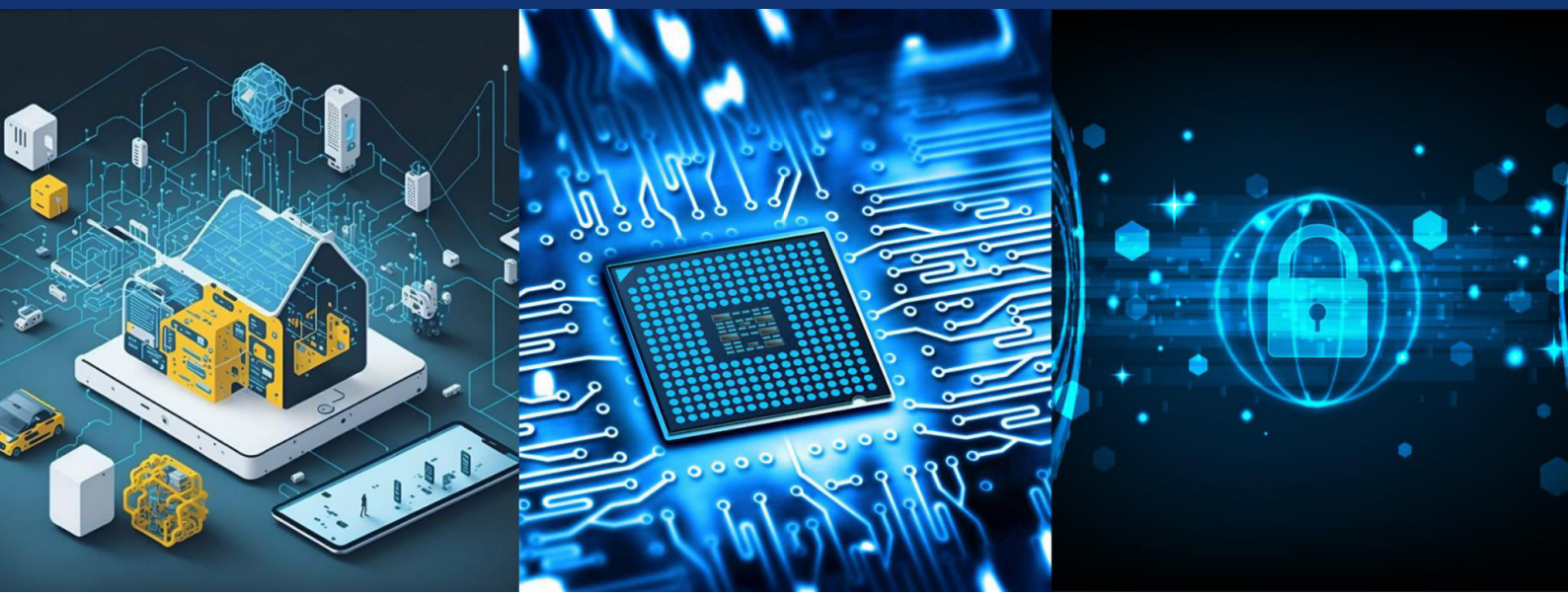
IJIRCCE

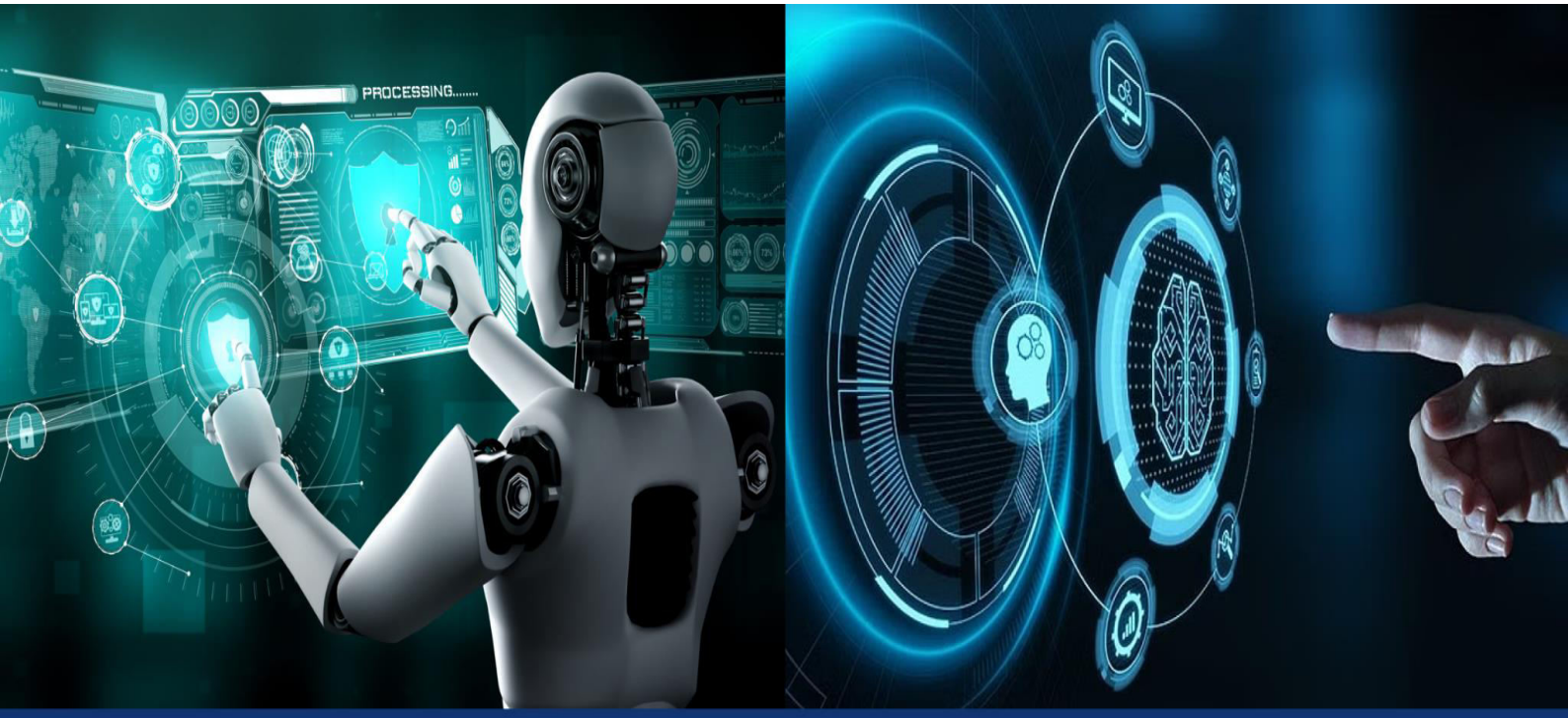
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 5, May 2025

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305185

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Sentiment Analysis of Social Media Presence using Machine Learning

Vishnu K[1], Pavan M[1], Darshan SR[1], Punith GR[1], Kaushik[1], Santhosh Kumar K L[2]

School of Computer Science & Engineering, Presidency University, Bengaluru, India[1]

Associate Professor, School of Computer Science and Engineering, Presidency University, Bengaluru, India[2]

ABSTRACT: This paper presents the Sentiment Analysis Using Machine Learning, a web-based tool designed for real-time sentiment analysis of Reddit comments. Leveraging pre-trained natural language processing models and the Reddit API, the system enables users to retrieve and analyze public sentiment on various topics or subreddits. The application utilizes efficient text preprocessing techniques and sentiment classification algorithms optimized for social media text to classify sentiments into positive, negative, or neutral categories. Visualizations such as pie charts, bar graphs, and word clouds facilitate intuitive understanding of sentiment trends. The modular design employs a lightweight Python backend and a Streamlit front-end interface, enabling easy deployment and interaction without extensive computational resources or model training. This tool supports users including researchers, marketers, and social scientists in gaining rapid insights into public opinions on social media platforms.

KEYWORDS: Sentiment Analysis, Reddit, Natural Language Processing, Social Media Mining, Pre-trained Models, Visualization

I. INTRODUCTION

In recent years, social media platforms have become prominent sources of public opinion, information dissemination, and social interaction. Among these platforms, Reddit stands out as a diverse and dynamic community-driven network with millions of users discussing a wide range of topics through posts and comments. Due to its topic-specific subreddits and threaded discussions, Reddit offers a rich and nuanced dataset for sentiment analysis, which is the computational study of opinions, emotions, and attitudes expressed in text. Extracting sentiment from social media content has become essential for applications such as market research, public opinion tracking, brand reputation management, and political analysis.

Sentiment analysis involves automatically identifying and categorizing subjective information within text data, commonly classifying sentiments as positive, negative, or neutral. Traditional sentiment analysis systems often focus on Twitter or Facebook data due to their popularity and accessibility. However, Reddit's unique structure and conversational style pose distinct challenges and opportunities. The volume and diversity of Reddit comments necessitate robust tools capable of efficiently analyzing large datasets while maintaining accuracy and interpretability. This paper aims to address this need by developing an accessible, real-time sentiment analysis tool specifically tailored for Reddit comments. Unlike systems that require extensive model training or computational resources, our solution lever- ages pre-trained, lexicon-based natural language processing (NLP) models such as VADER and TextBlob. These models are well-suited for social media text due to their sensitivity to emotive language, slang, and informal expressions, making them effective for classifying sentiment in Reddit's conversa- tional context.

Key to the system's design is the integration with Reddit's API using the Python Reddit API Wrapper (PRAW), which enables seamless, real-time data acquisition. Users can input keywords, topics, or subreddit names to fetch recent comments relevant to their interests. The retrieved data undergoes preprocessing steps, including noise removal, tokenization, and normalization, to prepare text for sentiment classification. The classified results are then presented through an intuitive, interactive graphical user interface (GUI) built with Streamlit, facilitating easy exploration of sentiment trends, distribution, and key textual insights via visualizations such as bar charts, pie charts, and word clouds. The modular architecture ensures the system is scalable and adaptable for various user groups, including researchers, marketers, social scientists, and casual users interested in public opinion mining. This democratizes sentiment analysis by reducing technical barriers and providing immediate access to insights derived from social media conversations.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

While the use of pre-trained models enables rapid deployment, it also means the system can be further enhanced by integrating fine- tuned deep learning models or multi-lingual support in future iterations. Overall, the Reddit Sentiment Explorer contributes to the growing field of social media analytics by offering a lightweight, user-friendly, and effective solution for real-time sentiment analysis on Reddit. The project demonstrates how existing NLP tools can be combined with modern web frame- works to build practical applications that address real-world needs in monitoring and understanding public sentiment on diverse social platforms.

II. RELATED WORK

Hutto and Gilbert [1] introduced VADER, a lightweight rule-based sentiment analysis tool optimized for social media. Utilizing a 7,500-feature lexicon with heuristics for emphasis and negation, VADER gained popularity for its adaptability to informal text. Despite its strengths, it cannot learn from new data or interpret sarcasm effectively.

Zhang and Wallace [2] compared deep learning architectures for sentiment analysis. Their study revealed CNNs extract features efficiently from short texts, while LSTMs capture temporal dependencies in medium-length sequences. Transformer models demonstrated superior context awareness but required significant computational resources and large datasets, presenting challenges in practical implementation.

Loria [3] developed TextBlob, a Python library offering accessible NLP tools including sentiment analysis. It calculates polarity and subjectivity using lexicon-based approaches, making it ideal for rapid prototyping. However, TextBlob struggles with complex language and cannot adapt dynamically to new contexts or cultural nuances.

Liu [4] provided a cornerstone survey categorizing sentiment analysis into document-level, sentence-level, and aspect-level approaches. The paper explored machine learning and lexicon-based techniques, highlighting feature selection methods. Despite advances, challenges persist in implicit sentiment detection, co-reference resolution, and domain dependency across sentiment analysis applications.

Agarwal et al. [5] developed a hybrid approach for Twitter sentiment classification, combining syntactic and semantic feature engineering with machine learning. Their framework utilized POS tagging and tree kernel methods to capture structural relationships in brief, noisy microblog content, pioneering solutions for platform-specific sentiment challenges.

Mohammad and Turney [6] examined text preprocessing effects on sentiment classification. Their experiments revealed no universal best-practice pipeline exists—preprocessing must be data-specific. Operations like stemming and stopword removal impact performance differently across datasets, empha- sizing the importance of intentional preprocessing design in sentiment analysis systems.

Singh and Sharma [7] presented a real-time sentiment analysis platform for businesses using Flask and TextBlob. Their system integrated front-end collection with back-end analysis and visualization, demonstrating accessible sentiment analytics without complex infrastructure. Primary limitations included weak contextual understanding and lack of multilingual support.

Deshmukh and Kale [8] applied VADER to YouTube comment sections to evaluate viewer sentiment. Their methodology included processing comment data via the YouTube API and visualizing sentiment distributions with Matplotlib. The re- search highlighted significant challenges in handling unstructured internet language and evolving online communication patterns.

Pontiki et al. [9] focused on aspect-based sentiment analysis using deep learning with attention mechanisms. Their approach assigned sentiment scores to specific product features rather than entire documents. Despite innovation, ABSA faces significant challenges in data availability, implicit aspect detection, and cross-domain generalization.

Medhat et al. [10] evaluated traditional machine learning classifiers for sentiment analysis. Their study showed SVM achieved best accuracy while Naive Bayes offered fastest processing. The research highlighted fundamental limitations in contextual awareness and domain adaptation for traditional approaches, suggesting the need for more advanced techniques.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305185

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Modern sentiment analysis must balance accuracy, efficiency, and adaptability. Rule-based approaches offer practical solutions for real-time applications, while deep learning provides superior context understanding with higher computational costs. Hybrid approaches combining rules with statistical or neural models represent promising directions for robust sentiment analysis systems.

The literature demonstrates evolution from lexicon-based methods to sophisticated neural architectures. Addressing persistent challenges in sarcasm detection, multilingual support, and aspect-level analysis remains crucial for next-generation sentiment systems that deliver meaningful insights across diverse applications and domains.

III. METHODOLOGY

This project is designed to provide real-time sentiment analysis of Reddit comments using pre-trained NLP models, Reddit API integration, and visual analytics. This section elaborates on the modular workflow, highlighting data acquisition, preprocessing, sentiment classification, and user interface rendering. The following architecture was implemented to support modularity, scalability, and ease of use.

A. System Architecture Overview

The system uses user-item interaction data and transforms it into a graph structure. Users and products are represented as nodes and interactions (ratings) are represented as weighted edges. The architecture involves the following stages.

- User Interface
- Reddit API Integration (via PRAW)
- Data Preprocessing Module
- Sentiment Classification Engine (TextBlob VADER)
- Visualization and Analytics Module

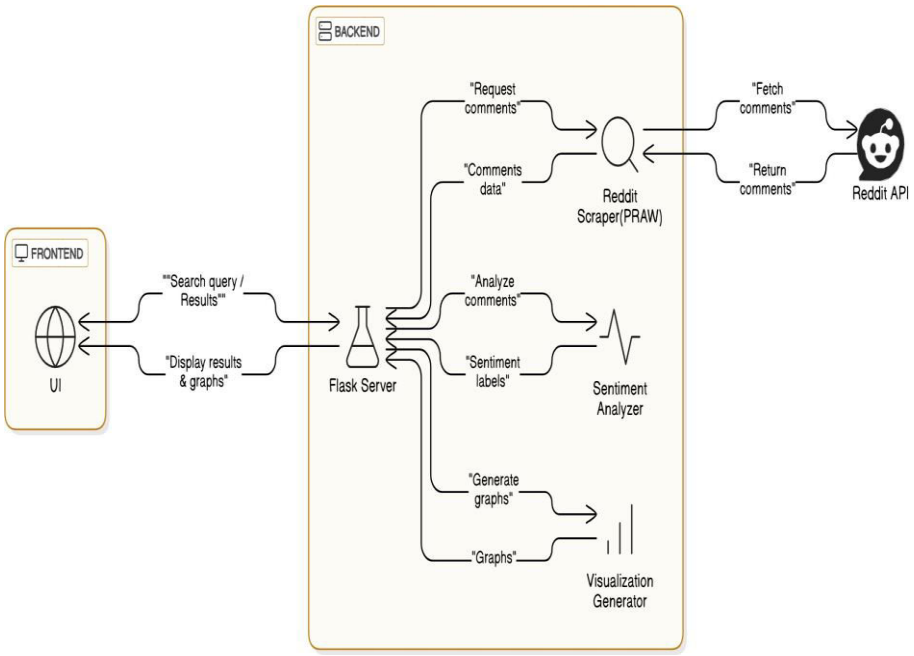


Fig. 1. System Architecture of the GNN-based Recommender

B. Data Acquisition

Reddit data is acquired in real-time using the Python Reddit API Wrapper (PRAW). Based on the user's input—such as a subreddit name, keyword, or post ID—the system fetches the most recent comments or submissions.

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305185

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- API Endpoint Access: OAuth 2.0 authentication is used via a Reddit developer account to ensure secure access.
- Query Scope: Users can specify filters such as top posts, comment limits, and date ranges to target specific datasets.

C. Text Preprocessing

Preprocessing is essential to clean and normalize raw text data before feeding it into sentiment classifiers. The following steps are performed:

- Lowercasing all text to maintain consistency
- Removal of punctuation and non-alphabetic characters
- Stopword removal using NLTK
- Tokenization into individual words
- Lemmatization for reducing words to their root form

D. Sentiment Classification

Two lexicon-based sentiment analysis models are integrated:

- **TextBlob:** Computes polarity and subjectivity of text. It is fast and effective for basic sentiment scoring.
- **VADER (Valence Aware Dictionary for sentiment Reasoning):** Optimized for social media texts, handling emojis, slang, punctuation, and capitalization.

Each comment is evaluated and labeled as Positive, Negative, or Neutral based on sentiment polarity.

E. Visualization and Insights

Visualizations are generated using matplotlib, seaborn, and wordcloud. These include:

- Pie or bar charts showing sentiment breakdown
- Lists of top positive and negative comments
- Word clouds for frequent terms per sentiment
- Optional trend plots for time-based sentiment tracking

F. Interface and Usability

The front-end is built with Streamlit, providing a responsive and interactive dashboard:

- Input fields for topic, subreddit, or keyword
- Button controls to trigger analysis
- Output panels for sentiment labels, stats, and charts

IV. IMPLEMENTATION DETAILS

The Reddit Sentiment Explorer was implemented using Python and Streamlit for the front-end, with NLP libraries including TextBlob and VADER for sentiment classification. The system integrates with Reddit via the Python Reddit API Wrapper (PRAW), allowing real-time comment retrieval and analysis.

A. Backend Components

- **Reddit API Integration:** Implemented using PRAW, with authentication to securely fetch live data from user-specified subreddits or post IDs.
- **Preprocessing Pipeline:** Utilized NLTK for tokenization, stopword removal, and lemmatization. Regex expressions were used to clean noisy text inputs.
- **Sentiment Engine:** VADER and TextBlob were used as pre-trained sentiment analyzers. Comments are passed through both for comparative labeling and scoring.
-

B. Frontend (GUI)

The graphical user interface was developed using HTML, CSS and Js to ensure interactivity and ease of access. Key interface features include:

- Text boxes for subreddit/topic input
- Sidebar filters for comment limits and sorting preferences
- Real-time display of sentiment analysis results
- Auto-generated visualizations after processing

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305185

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

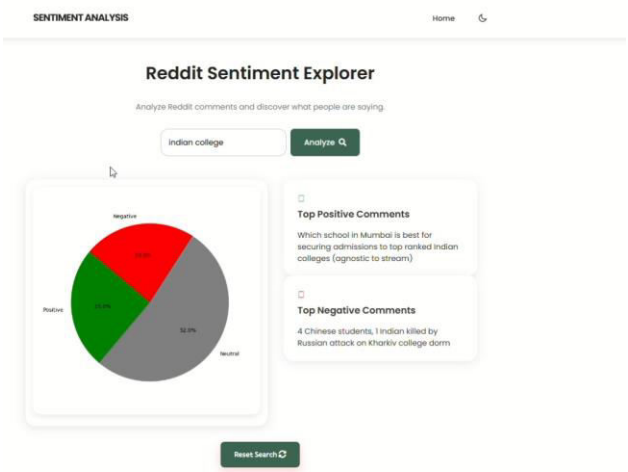


Fig 2. Frontend (GUI)

V. RESULTS AND VISUALIZATION

The system was tested on a range of subreddits including r/technology, r/worldnews, and r/movies. For each test, a fixed number of recent comments were fetched, processed, and categorized. Below are the sample results and insights.

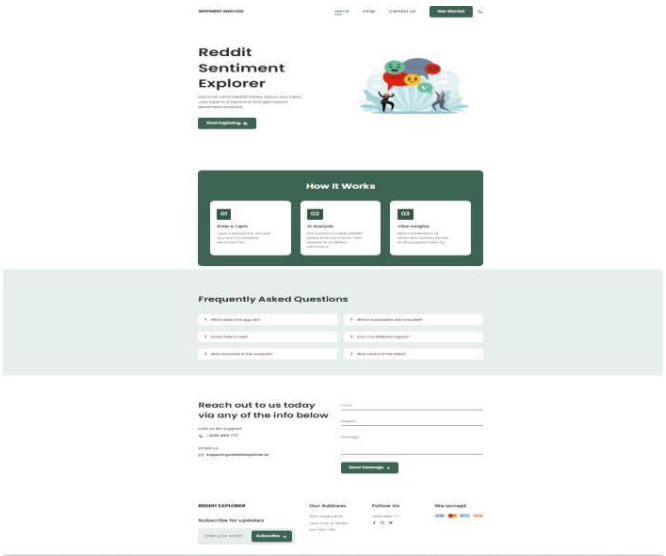


Fig. 3. Sentiment Distribution for Sample Reddit Data

A. Sentiment Distribution

A summary of sentiment distribution (positive, neutral, negative) for 200 comments in r/technology is shown in Fig. 3.

B. Sample Comments

Below are top-scoring examples from the sentiment analysis:

- **Positive:** "This update is a game-changer. Kudos to the developers!"
- **Negative:** "Absolutely terrible experience. It's not usable anymore."
- **Neutral:** "I read the article. Interesting points mentioned."

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. System Responsiveness

The system typically takes under 3 seconds to process and analyze 100 comments. Average response time was measured across 10 runs and found to be 2.8 seconds per batch.

D. User Testing and Feedback

Informal testing was conducted with 10 users, who high- lighted:

- Ease of use and clarity of sentiment outputs
- Desire for more advanced search filters (to be included in future versions)

VI. CONCLUSION

In this paper, we presented the design and development of Reddit Sentiment Explorer, a sentiment analysis system tailored to extract, process, and interpret public sentiment from Reddit posts and comments. This project bridges the gap between textual data mining and intuitive user-level insight generation, offering a streamlined interface for ex- ploring sentiments across Reddit communities. By leveraging pre-trained sentiment analysis models and efficient natural language processing pipelines, our system enables real-time sentiment classification without the need for model training, making it suitable for lightweight and rapid deployment.The project employs fundamental NLP techniques includ- ing tokenization, lemmatization, TF-IDF vectorization, and lexicon-based scoring to classify user-generated content into positive, negative, or neutral sentiments. The integration of Streamlit for GUI development allowed us to create an in- teractive and accessible platform for users to view sentiment trends, top comments, and overall post polarity in a visual format. This adds significant value for researchers, marketers, and social media analysts looking to monitor public opinion. Our findings emphasize the utility of using pre-existing NLP tools such as VADER and TextBlob in domain-specific contexts like Reddit. While the current system does not in- volve training custom models, it remains extensible for future development, including incorporation of deep learning models and multilingual support.Future work may include expanding data sources beyond Reddit, adding advanced visual analytics, and building adap- tive learning pipelines to handle sentiment drift. Overall, Reddit Sentiment Explorer demonstrates the potential for combining NLP, data visualization, and user-centric design to extract actionable insights from social discourse platforms.

REFERENCES

1.J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 8, no. 1, 2014, pp. 216–225.

2.Y. Zhang and B. D. Wallace, “A Sensitivity Analysis of (and Practi- tioners’ Guide to) Convolutional Neural Networks for Sentence Classi- fication,” in *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, 2015, pp. 253–263.

3.S. Loria, “TextBlob: Simplified Text Processing,” [Online]. Available: https://textblob.readthedocs.io/

4.Liu, “Sentiment Analysis and Opinion Mining,” *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

5.Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment Analysis of Twitter Data,” in *Proceedings of the Workshop on Languages in Social Media*, 2011, pp. 30–38.

6.S. M. Mohammad and P. D. Turney, “NRC Emotion Lexicon,” *Na- tional Research Council Canada*, 2013. [Online]. Available: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

7.S. Singh and R. Sharma, “Real-time Sentiment Analysis System Using Flask and TextBlob,” *International Journal of Computer Applications*, vol. 180, no. 35, pp. 25–29, 2018.

8.P. Deshmukh and M. Kale, “Sentiment Analysis on YouTube Com- ments Using VADER,” in *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 871–875.

9.M. Pontiki et al., “SemEval-2016 Task 5: Aspect Based Sentiment Analysis,” in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, 2016, pp. 19–30.

10.W. Medhat, A. Hassan, and H. Korashy, “Sentiment Analysis Algorithms and Applications: A Survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details