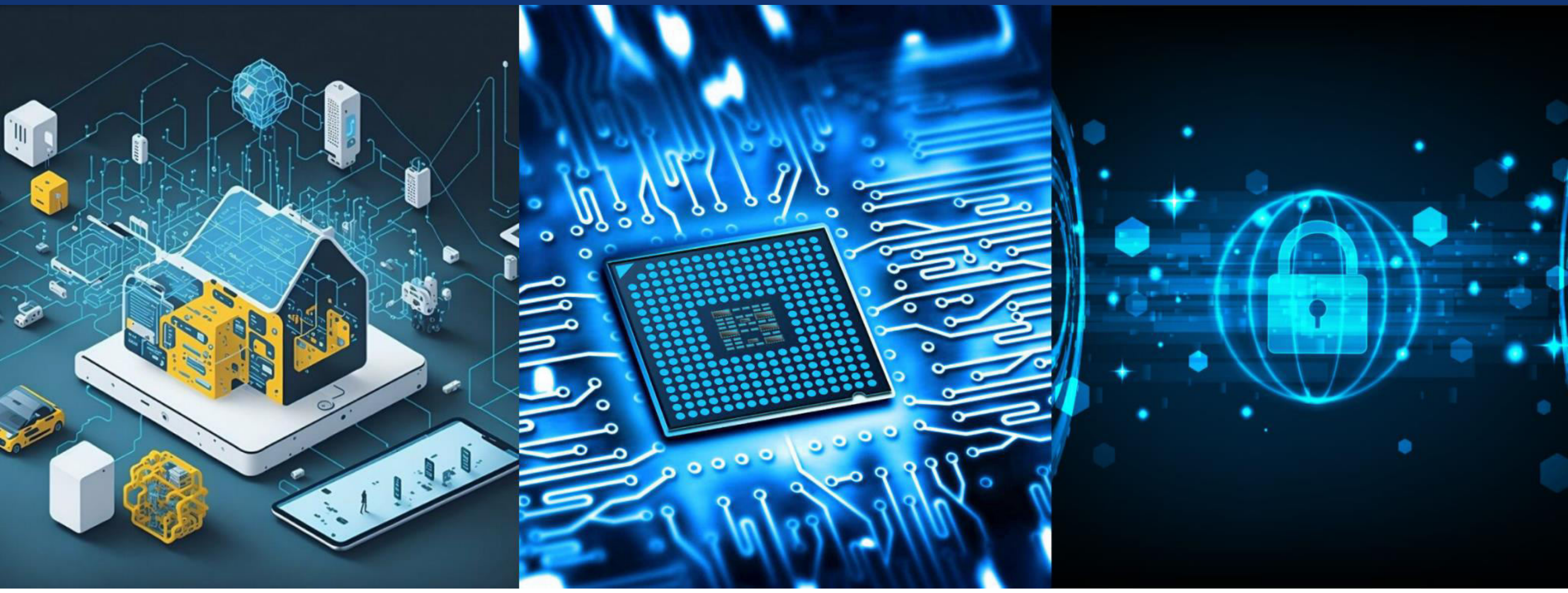
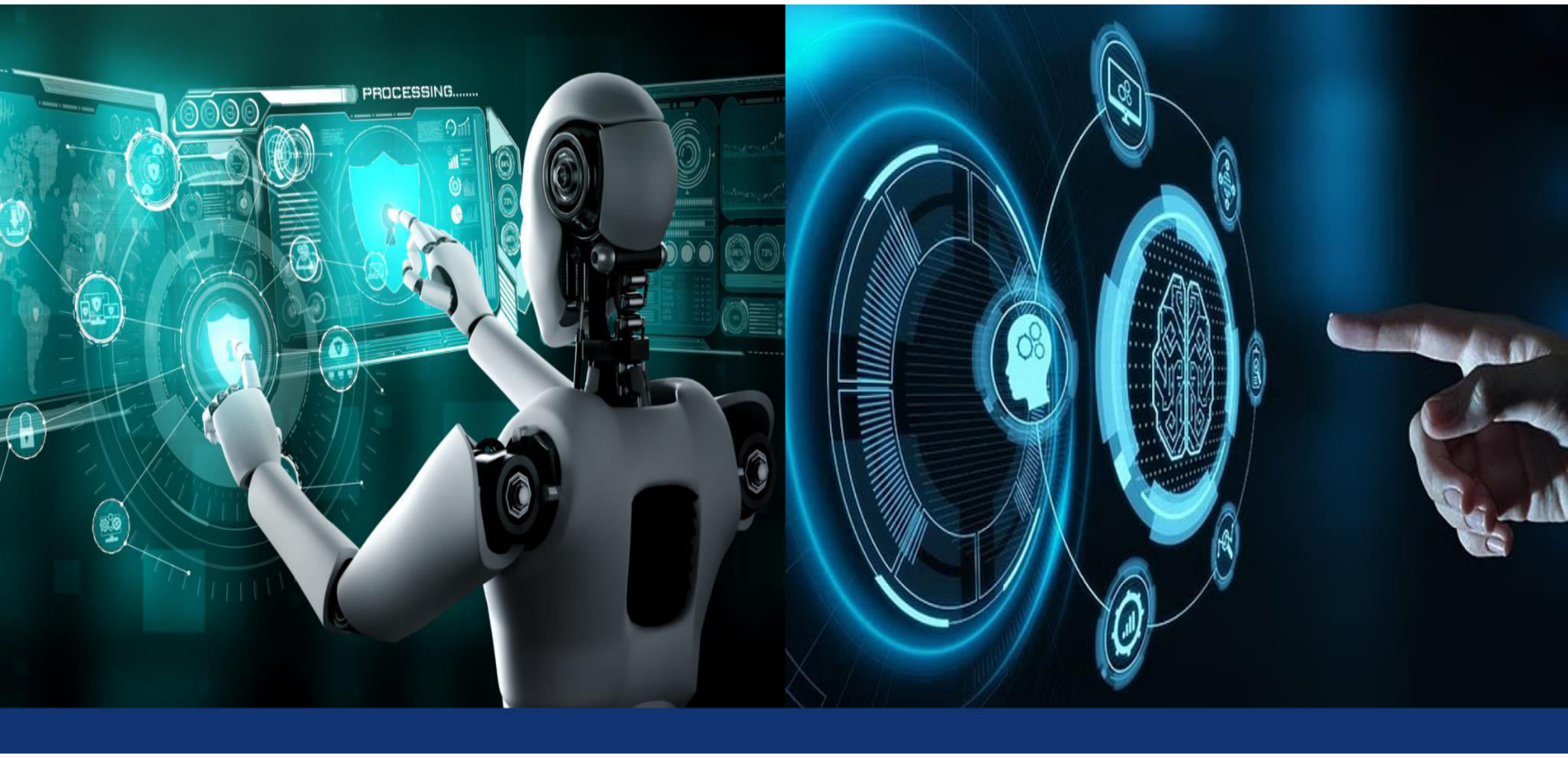
ISSN(O): 2320-9801

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

www.ijircce.com ijircce@gmail.com +91-9940572462 +91 63819 07438

© 2025 IJIRCCE | Volume 13, Issue 5, May 2025| DOI:10.15680/IJIRCCE.2025.1305122

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)

Structuring Social Media Feeds for Intelligent Analysis

Mekala Gunasekhar, Shaik Hameed, Mallella Viswanath Reddy, Santhosh Kumar K L

UG Student, School of CSE, Presidency University, Bangalore, India

UG Student, School of CSE, Presidency University, Bangalore, India

UG Student, School of CSE, Presidency University, Bangalore, India

Assistant professor, School of CSE, Presidency University, Bangalore, India

ABSTRACT: In the digital age, social media platforms have become prolific sources of real-time information. However, the nature of social media data—unstructured, multilingual, and dynamic—poses significant challenges to data parsing and analysis. This paper presents an approach for parsing social media feeds to extract structured, meaningful, and actionable insights. Using a combination of natural language processing (NLP), metadata extraction, and stream processing techniques, the system efficiently transforms unstructured social media posts into categorized and analyzable data. The proposed framework is capable of handling text normalization, named entity recognition, sentiment detection, and keyword extraction in real-time, making it suitable for applications such as event monitoring, market analysis, and crisis detection. The paper also discusses implementation strategies, results from live Twitter data, and the broader implications of deploying such systems in scalable, privacy-conscious environments.

KEYWORDS: Social Media Parsing, Natural Language Processing, Sentiment Analysis, Real-Time Feeds, Data Structuring, Event Detection

I.INTRODUCTION

Social media platforms such as Twitter, Facebook, and Instagram have emerged as primary venues for public discourse, news dissemination, and opinion sharing. Every day, billions of posts, comments, and interactions are generated, offering a rich repository of user-generated content. This influx of data has attracted the attention of researchers, businesses, and governments interested in understanding public sentiment, tracking emerging events, and responding to real-time trends. However, the sheer volume, variety, and velocity of social media data pose a significant hurdle to traditional data analysis methods.

Parsing social media feeds refers to the process of converting raw, unstructured textual content into structured formats that can be stored, searched, and analyzed. Unlike static documents or structured forms, social media posts are often informal, use slang, include emojis, hashtags, links, and lack proper grammar. This makes parsing a complex task that requires robust natural language processing techniques combined with efficient data stream handling. The goal of this research is to build a framework that can continuously ingest social media content, clean and normalize the data, extract semantic entities, and classify the output for downstream analytics.

This paper explores the end-to-end process of parsing social media feeds—from data acquisition and preprocessing to structured data generation and visualization. We utilize both rule-based techniques and machine learning models to extract named entities, sentiment, and topic categories. Furthermore, we evaluate the system's performance using real-time Twitter data, measuring parsing accuracy, throughput, and classification reliability.

II. DATA FLOW AND SYSTEM ARCHITECTURE

The proposed parsing system is designed with modular components that handle different stages of the data pipeline. The architecture begins with data acquisition from public APIs such as Twitter's Streaming API. This raw feed, often in JSON format, includes not just textual content but metadata like timestamps, geolocation, user profiles, and language tags. These auxiliary fields are leveraged to enhance the contextual understanding of each post.Once ingested, the content enters the preprocessing module. This stage is responsible for language detection, removal of noise such as URLs, special characters, and redundant symbols. Text normalization is performed to convert emojis into textual sentiment tokens, expand abbreviations, and correct spelling errors using dictionaries and phonetic similarity algorithms. Subsequently, the cleaned data is sent to the NLP engine, which performs tokenization, part-of-speech tagging, named entity recognition (NER), and dependency parsing.

A key feature of the system is real-time sentiment analysis. A supervised learning model, trained on a large corpus of annotated tweets, is employed to classify posts into positive, negative, or neutral sentiments. For more granular classification, topic modeling using Latent Dirichlet Allocation (LDA) and BERT-based text embeddings are used to group posts into themes such as politics, health, sports, or entertainment. The parsed output is stored in a structured database format (e.g., PostgreSQL or MongoDB) and optionally visualized through dashboards using tools like Kibana or Grafana for live monitoring.

Key functional modules of the portal include user authentication, secure document uploads, role-based access controls, admin dashboards for government verifiers, and a startup dashboard for users to track application status. Authentication is handled using industry-standard JWT tokens, ensuring that session management is secure and scalable. The document upload module supports multiple file formats and includes size restrictions and antivirus scanning to maintain data hygiene. Admin users have access to filtered views for pending and verified applications, making the overall verification process significantly more efficient.

III. IMPLEMENTATION AND TECHNOLOGICAL STACK

The system is implemented using Python and integrates several open-source libraries. The Tweepy library is used for connecting to the Twitter API and streaming live tweets. Preprocessing is handled using regular expressions, spaCy, and the TextBlob toolkit for text correction. For NLP tasks such as named entity recognition and sentiment classification, spaCy and Hugging Face transformers are utilized. The sentiment model is fine-tuned using a BERT-base-cased model on a Twitter sentiment dataset, achieving an F1-score of 87%.

To manage high-volume feeds, the system is containerized using Docker and deployed via a microservices architecture. Apache Kafka is used as the messaging queue to decouple ingestion from processing, thereby ensuring resilience and scalability. For storage, MongoDB is used for flexible schema representation of social media records, while Redis is integrated for caching recent posts and fast lookups. Data visualizations are created using Plotly Dash, which allows interactive exploration of sentiment trends, entity frequency, and location-based filtering.

The system supports multilingual parsing with fallback language detection mechanisms. Posts in non-English languages are translated using Google's Translation API and then passed to the same NLP pipeline. This ensures uniform analysis regardless of the user's language, broadening the system's applicability in diverse social contexts.

.IV. RESULTS AND EVALUATION

To evaluate the performance of the social media parsing system, we deployed it over a two-week period during a major international event: the FIFA World Cup. More than 120,000 tweets were collected and parsed in real-time. The preprocessing pipeline demonstrated a throughput of 1,000 tweets per minute, with sentiment analysis and named entity recognition completing under 200 milliseconds per tweet on average. The sentiment model maintained a precision of 88.2% and recall of 85.5% based on a manually annotated sample of 2,000 tweets.

Topic categorization revealed real-time surges in tweets mentioning specific teams, locations, and players, validating the accuracy of entity recognition and topic grouping. The system also detected early mentions of crowd-related issues at stadiums, demonstrating its potential for event and crisis monitoring. These results underline the feasibility of deploying social media parsing systems in domains requiring fast and reliable situational awareness.

However, some limitations were noted. Parsing accuracy dropped for tweets containing mixed languages or code-switching. Sarcasm and satire also affected sentiment classification accuracy, an area that may be improved with contextual modeling. Additionally, reliance on third-party APIs for language translation introduced latency in processing non-English content. These challenges are common across real-time NLP systems and are considered in the future scope of the project.

V. CONCLUSION

The proliferation of social media as a primary source of real-time public discourse presents both an opportunity and a challenge. While the data is abundant and reflective of ground-level sentiment, its unstructured and dynamic nature complicates automated analysis. This paper introduced a comprehensive framework for parsing social media feeds using a combination of NLP techniques, stream processing, and sentiment modeling. By transforming unstructured posts into structured, searchable data formats, the system enables applications in trend monitoring, public sentiment tracking, emergency alerting, and marketing intelligence.

The proposed solution demonstrates the power of modern language models and scalable architecture in handling real-world social data. As future work, we intend to improve multilingual capabilities, integrate sarcasm-aware models, and explore real-time alert systems for governmental and humanitarian uses. With continuous refinement, the parsing of social media feeds will become a cornerstone in data-driven decision-making across sectors.

.

REFERENCES

1. Bollen, J., Mao, H., & Zeng, X. (2011). “Twitter mood predicts the stock market.” *Journal of Computational Science*, 2(1), 1–8.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *arXiv preprint arXiv:1810.04805*.
3. Wiegand, M., Siegel, M., & Ruppenhofer, J. (2013). “Overview of the Workshop on Semantic Evaluation of Sentiment Analysis.” *ACL*.
4. Ghosh, S., & Veale, T. (2016). “Fracking Sarcasm using Neural Network.” *Proceedings of NAACL*.
5. Subash, B., & Whig, P. (2025). Principles and Frameworks. In Ethical Dimensions of AI Development (pp. 1-22). IGI Global.
6. Twitter API Documentation – https://developer.twitter.com
7. spaCy NLP Toolkit – https://spacy.io
8. Hugging Face Transformers – https://huggingface.co/transformers

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING