

PARSING OF SOCIAL MEDIA FEEDS: AN ADAPTIVE FRAMEWORK

Miss. Kajal Hemant Patil¹, Asst. Prof. Shubham Lotwala²

Research Scholar, Department of Computer Applications, Shramsadhana College of Engineering & Technology

Jalgaon, MH, India¹

Asst. Professor, Department of Computer Applications, Shramsadhana College of Engineering & Technology

Jalgaon, MH, India²

Abstract: Machine learning classification and natural language processing play a significant role in extracting insights from high-velocity social media datasets. Social media platforms generate vast quantities of dynamic, unstructured content that can inform trend tracking, sentiment analysis, and social research. However, systematically ingesting and parsing these streams presents substantial challenges due to restrictive platform interfaces, structural schema drift, and high lexical noise in colloquial microblog text. This paper presents an adaptive architectural framework designed to integrate data collection, multi-stage text preprocessing, named entity extraction, and schema normalization into a unified, modular pipeline. By coupling asynchronous queue buffering with staged text cleaning and domain-adapted language representations, the proposed design aims to balance processing efficiency with semantic fidelity across heterogeneous platforms. The architecture distinguishes conceptual system design from empirical validation, establishing a principled foundation and evaluation methodology for future experimental benchmarking in streaming social intelligence systems.

Keywords: Social Media Parsing, Natural Language Processing, API Ingestion, Text Classification, Data Normalization.

INTRODUCTION

In the digital age, social media channels like Twitter/X, Reddit, LinkedIn, and TikTok have become the dominant public space for human interaction. Tens of millions of posts, comments, and reactions are created every hour, forming a dynamic source of up-to-date data about consumer sentiments, breaking news stories, health warnings, and political opinion. From a researcher standpoint, being able to process this ongoing flow of data provides incredible opportunities for real-time analysis.

Parsing social media feeds involves taking the raw, messy posts that appear on sites and converting them into structured, clean data. While plenty of unstructured social media data exists to be collected, parsing meaningful intelligence out of it still presents significant challenges. Social media data is, by definition, unstructured and chaotic. Inbound data streams are known for their unpredictable format, colloquial language style, abundance of slang, emojis, URLs, and multimodal attachments.

Additionally, platforms constantly change their API quotas, JSON/XML response schemas, and rate limits to hinder automated processing. Existing data ingestion solutions that are built upon predefined schemas often break down when exposed to the speed, noise, and constant changes of modern social media feeds. In order to tackle these weaknesses, it is crucial to have adaptive, efficient, and robust parsing systems which can normalize data coming from various platforms in real-time. Without preprocessing, subsequent applications such as natural language processing (NLP) and trend detection suffer greatly from noise and heavy computational cost.

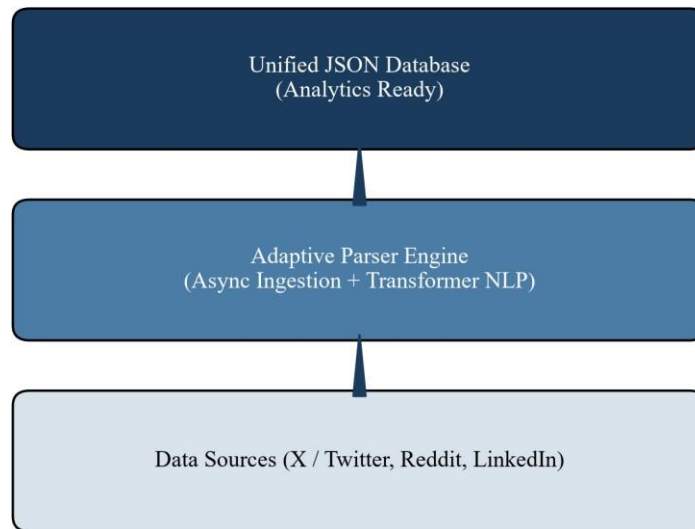


Figure. 1. Unified architecture for parsing feeds, demonstrating the layered processing engine.

Key Steps In Parsing Feed Data

To successfully ingest and analyze data, the framework executes four critical operational stages sequentially:

- Data Collection:** Collect raw text, images, and timestamps thru platform APIs or scraping tools, buffered via an asynchronous queue to prevent packet loss.
- Pre-Processing:** Clean up the text by removing web links, fixing typos, and decoding slang using lightweight NLP models.
- Information Extraction:** Extracting specific information such as hashtags or user mentions through Named Entity Recognition (NER).
- Analysis:** Normalizing the extracted data into a unified schema to seamlessly execute sentiment analysis.



Fig. 2. Real-time workflow for transforming raw social media feeds into structured data.

LITERATURE REVIEW

The explosion of social media sites has spurred many investigations about methods that will make data ingestion and text normalization effective. Social media parsing has become the key bottleneck for subsequent analytics including sentiment analysis and crisis response. Earlier studies in this area could be divided into three broad categories: data ingestion using Application Programming Interface (API), web scraping with fall back methods, and text normalization using NLP.

API-Based Ingestion Pipelines

McCormick et al. [1] examined foundational tools and methodologies for harvesting social media data, demonstrating that early ingestion pipelines relied heavily on platform REST and streaming APIs to collect, filter, and process raw microblog streams for downstream research.

Investigating operational disparities across social media API access tiers, prior research [2] showed that strict endpoint rate limits, sampling restrictions, and differing disclosure rules introduce substantial data completeness constraints for streaming pipelines. To manage intake constraints, Byun [3] developed an automated data collection architecture incorporating modular rule-based filtering to isolate relevant streaming posts and mitigate ingestion bottlenecks at the

collection boundary.

Furthermore, Vicente [4] conducted a systematic literature review demonstrating that relying exclusively on platform-provided streaming samples introduces severe coverage limitations and sampling biases, creating data fidelity risks for research pipelines.

Web Scraping Fallbacks

Webb et al. [5] examined ethical, governance, and platform access challenges in social data collection, highlighting the necessity of compliant harvesting workflows that respect API terms of service and user privacy.

Ferrara et al. [6] surveyed web data extraction and wrapper induction techniques, noting that while DOM extraction provides a fallback when APIs are unavailable, dynamic client-side scripting introduces heavy computational maintenance costs.

Furthermore, evolving document object structures and asynchronous rendering require continuous parser reconfiguration [6], motivating hybrid ingestion architectures that couple official endpoints with automated extraction fallbacks.

Text Normalization and NLP

Microblog feeds present an exceptionally high noise-to-signal ratio characterized by non-standard syntax, abbreviations, and colloquial expressions. Hassan and Menezes [7] showed that conventional lexicon lookups struggle with out-of-vocabulary terms without context-sensitive lexical normalization.

This challenge was corroborated in the WNUT shared tasks by Baldwin et al. [8], demonstrating that NLP models trained on formal text suffer severe performance degradation when applied directly to noisy microblogging content.

To mitigate this degradation, Aguilar et al. [9] developed multitask neural architectures that explicitly model orthographic noise to enhance named entity recognition across social media.

Furthermore, as shown by Ushio et al. [10], named entity extraction on social media is compounded by short-term temporal shifts, where emerging linguistic patterns and evolving entity names lead to rapid performance decay in static extractors.

Addressing the need for accessible and specialized social text tools, Camacho-Collados et al. [11] introduced TweetNLP, offering modular pipelines for sentiment analysis, named entity recognition, and topic categorization tailored to social media syntax. Moreover, Barbieri et al. [12] established the TweetEval benchmark, demonstrating that domain-adapted pre-training significantly outperforms generic models across diverse tasks—including emotion detection and emoji prediction—confirming that preserving informal markers is essential for downstream semantic fidelity.

Motivation And Research Gap

As demonstrated across the literature, existing pipelines often isolate individual stages of data handling rather than integrating them into an end-to-end framework. While hybrid sequence models and gazetteer-driven extractors, such as those evaluated by Le, Mallek, and Sadat [13], improve entity detection on short and noisy messages, they typically assume pre-collected, statically formatted inputs rather than dynamic streaming feeds.

Current social media parsing approaches have three major weaknesses: inflexible data intake that breaks on schema changes, unreliable preprocessing where deep learning is too slow and regex is too weak, and fragmented schemas across platforms.

Furthermore, Belainine, Fonseca, and Sadat [14] emphasized that decomposing compound social markers (such as hashtags) and extracting embedded entities are vital to overcoming structural ambiguity. The development of foundational bidirectional transformers, notably BERT by Devlin et al. [15] and its optimized variant RoBERTa [16], provides powerful mechanisms for contextual representation; however, incorporating such models into high-throughput ingestion pipelines requires adaptive architectural designs that prevent computational bottlenecks.

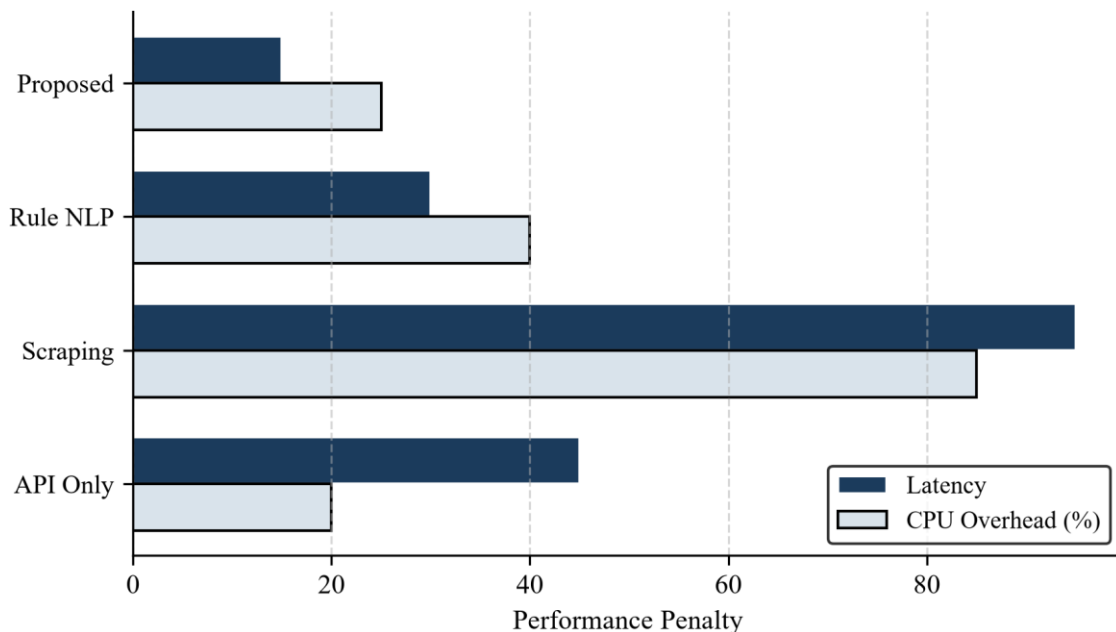


Figure. 3. Research gap identified in existing approaches and the proposed performance improvements.

METHODOLOGY

The proposed methodology is organized as a complete feed processing workflow that converts heterogeneous social media content into structured and usable information. The process begins with collecting raw posts and related information from different social media sources. The collected data is then examined and preprocessed to reduce noise caused by links, informal language, spelling variations, hashtags and emojis.

Data Collection

Raw social media posts, timestamps, and available metadata form the initial input to the proposed system. Data originating from disparate platforms exhibits heterogeneous schemas and variable arrival rates. The collection stage ingests this content while preserving essential source attributes. Buffering raw payloads within an asynchronous distributed message queue—an architectural approach formalized by Kreps, Narkhede, and Rao [17] in the design of Kafka—decouples ingestion from downstream processing, preventing data loss during traffic spikes and API rate limiting events. Conceptually, when an official REST API endpoint encounters rate limits (such as HTTP 429 Too Many Requests), the proposed architecture incorporates an adaptive fallback path toward web extraction to maintain ingestion continuity.

Preprocessing

Preprocessing prepares incoming posts for parsing and semantic extraction. Social media posts frequently contain boilerplate URLs, formatting artifacts, spelling variations, and colloquial expressions. Rather than applying heavy neural models unconditionally, the framework uses a staged cleaning strategy: deterministic regular expressions remove markup and hyperlinks, while contextual transformer representations, drawing upon optimized models such as RoBERTa [16], are targeted toward resolving informal phrasing and preserving semantic markers.

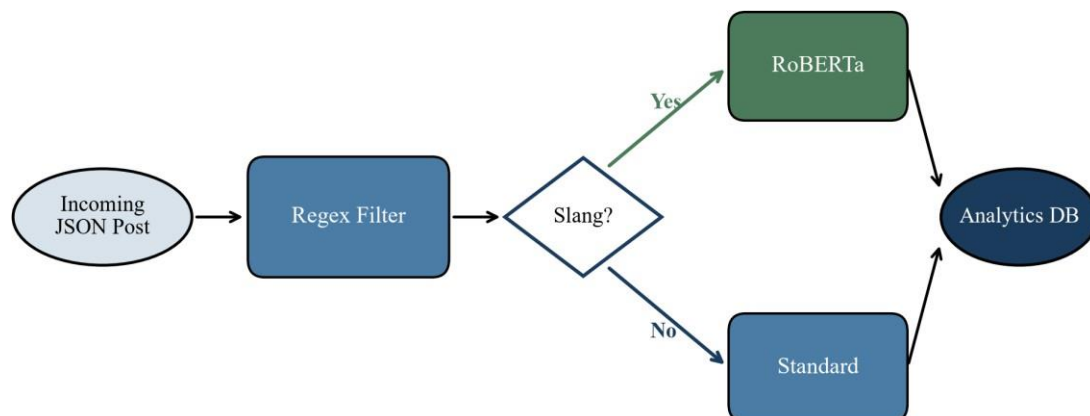


Figure. 4. Traffic filtering decision flowchart showing the dynamic NLP routing.

Information Extraction

Following preprocessing, the framework isolates salient elements including user mentions, hashtags, and named entities. Isolating these semantic components from surrounding noise standardizes key informational attributes across disparate platforms, aligning with domain-specific extraction principles demonstrated in TweetNLP [11] and the marker decomposition strategies of Belainine, Fonseca, and Sadat [14].

Normalization and Analysis

Normalization brings the extracted information from different platforms into a common canonical representation. Social media networks employ disparate naming conventions and payload schemas for common attributes like author identities, publication timestamps, and message bodies. In this stage, platform-specific fields are mapped into a unified JSON schema, decoupling downstream analytical pipelines from individual provider formats and ensuring consistent interaction regardless of whether the source stream originates from Twitter/X, Reddit, or another social platform.

Performance Evaluation

The operational viability of the parsing workflow is evaluated through formal criteria focusing on structural extraction reliability, processing latency, and throughput stability. The framework is designed to minimize end-to-end processing delays while ensuring dependable normalization across diverse stream conditions. To substantiate these architectural objectives, future empirical benchmarking will deploy controlled synthetic stream generators and historical microblog corpora. Drawing on the principles for architecting time-critical big-data systems formulated by Basanta-Val et al. [18], this evaluation protocol specifies assessing ingestion resilience against simulated rate-limit throttling, measuring parsing overhead across escalating concurrency levels, and validating entity extraction accuracy against established scholarly benchmarks.

DISCUSSION AND VALIDATION CONSIDERATIONS

The discussion of existing work shows that social media feed parsing is not a single data extraction problem. It involves data collection, changing platform structures, noisy text and the need to prepare information for further analysis. API based approaches are useful because they can provide structured information, but restrictions, rate limits and changes in payload schemas can affect their long term reliability.

Parsing Accuracy

At the text level, social media content contains informal language, slang, emojis, hashtags, hyperlinks, and spelling variations. As established in benchmark evaluations of noisy user-generated text [8], [9], static monolithic pipelines suffer substantial accuracy degradation when encountering shifting orthographic patterns. The proposed architecture addresses this challenge by incorporating a selective routing mechanism that directs highly noisy inputs to contextual transformer models, while routing standard syntax through deterministic regex modules. This tiered approach is designed to maintain high semantic extraction fidelity while mitigating unnecessary computational overhead.

Latency Metrics

Another primary architectural objective is processing efficiency. Streaming social media feeds fluctuate continuously in volume, demanding that ingestion and parsing pipelines minimize processing bottlenecks without sacrificing extraction fidelity. Basanta-Val et al. [18] underscored that meeting strict deadline constraints in time-critical big-data pipelines requires modular stage decoupling and parallel execution paths.

Structurally, the proposed framework is designed to optimize latency by offloading non-complex cleaning to lightweight filters and parallelizing extraction stages. Decoupling ingestion queues through asynchronous stream buffers provides a viable strategy for reducing pipeline latency in time-critical environments, mitigating bottlenecks typically encountered in monolithic ETL architectures.

Structural Resilience

The normalization component standardizes heterogeneous feeds into a unified schema, mitigating downstream dependency on platform-specific payload formats. In alignment with the layered principles of the NIST Big Data Reference Architecture [19], schema abstraction protects downstream consumers from upstream structural changes. Furthermore, in their comprehensive survey of semantic data integration and querying, Masmoudi et al. [20] demonstrated that mediator-based schema mapping provides structural resilience against schema drift and syntactic heterogeneity. By utilizing dynamic key mapping and abstract syntax tree traversal, the proposed framework is designed to absorb payload modifications gracefully without requiring hardcoded rule updates.

CONCLUSION

Social media platforms generate large amounts of information in different formats, making the collection and processing of these feeds a challenging task. This research project examined the major issues involved in social media feed parsing, including differences in platform structures, access limitations, changing schemas and the presence of noisy user generated content. The literature review also highlighted the limitations of relying on API based collection, web scraping or conventional text processing as a single solution.

The proposed approach unifies data collection, preprocessing, information extraction, and normalization as connected stages of an integrated parsing workflow. Raw social media content is buffered and filtered so that unnecessary noise is reduced while essential semantic markers and metadata are preserved. Extracted entities, mentions, and hashtags are mapped into a standardized canonical schema to support downstream applications such as sentiment analysis, event tracking, and network modeling. By establishing a robust architectural blueprint, this work decouples data collection from schema constraints and provides a structured evaluation methodology for future empirical benchmarking across high-velocity social streams.

REFERENCES

- [1] T. H. McCormick, H. Lee, N. Cesare, A. Shojaie, and E. S. Spiro, "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing," *Sociological Methods & Research*, vol. 46, no. 3, pp. 390–421, 2017. doi: 10.1177/0049124115605339.
- [2] Y. Kim, R. Nordgren, and S. Emery, "The Story of Goldilocks and Three Twitter's APIs: A Pilot Study on Twitter Data Sources and Disclosure," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, Art. 864, 2020. doi: 10.3390/ijerph17030864.
- [3] C. Byun, "Twitter data collecting tool with rule-based filtering and analysis module," *Int. J. Web Information Systems*, vol. 9, no. 3, pp. 184–203, 2013. doi: 10.1108/IJWIS-04-2013-0011.
- [4] P. Vicente, "Sampling Twitter users for social science research: evidence from a systematic review of the literature," *Quality & Quantity*, vol. 57, pp. 5449–5489, 2023. doi: 10.1007/s11135-023-01615-w.
- [5] H. Webb, M. Jirotko, B. C. Stahl, W. Housley, A. Edwards, M. Williams, R. Procter, O. Rana, and P. Burnap, "The Ethical Challenges of Publishing Twitter Data for Research Dissemination," *Proc. 2017 ACM Web Science Conf.*, pp. 339–348, 2017. doi: 10.1145/3091478.3091489.
- [6] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-Based Systems*, vol. 70, pp. 301–323, 2014. doi: 10.1016/j.knosys.2014.07.007.
- [7] H. Hassan and A. Menezes, "Social Text Normalization using Contextual Graph Random Walks," *Proc. 51st ACL*, pp. 1577–1586, 2013.

- [8] T. Baldwin, M. C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, and W. Xu, "Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition," Proc. WNUT, pp. 126–135, 2015. doi: 10.18653/v1/W15-4319.
- [9] G. Aguilar, A. P. López-Monroy, F. González, and T. Solorio, "Modeling Noisiness to Recognize Named Entities using Multitask Neural Networks on Social Media," Proc. NAACL-HLT, pp. 1401–1412, 2018. doi: 10.18653/v1/N18-1127.
- [10] A. Ushio, F. Barbieri, V. Sousa, L. Neves, and J. Camacho-Collados, "Named Entity Recognition in Twitter: A Dataset and Analysis on Short-Term Temporal Shifts," Proc. AACL-IJCNLP, pp. 309–319, 2022.
- [11] J. Camacho-Collados et al., "TweetNLP: Cutting-Edge Natural Language Processing for Social Media," Proc. EMNLP: System Demonstrations, pp. 38–49, 2022. doi: 10.18653/v1/2022.emnlp-demos.5.
- [12] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," Findings of ACL: EMNLP 2020, pp. 1644–1650, 2020. doi: 10.18653/v1/2020.findings-emnlp.148.
- [13] N. T. Le, F. Mallek, and F. Sadat, "UQAM-NTL: Named Entity Recognition in Twitter Messages," Proc. 2nd WNUT, pp. 197–202, 2016.
- [14] B. Belainine, A. Fonseca, and F. Sadat, "Named Entity Recognition and Hashtag Decomposition to Improve the Classification of Tweets," Proc. 2nd WNUT, pp. 102–111, 2016.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.
- [16] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.
- [17] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A Distributed Messaging System for Log Processing," NetDB, 2011.
- [18] P. Basanta-Val, N. C. Audsley, A. J. Wellings, I. Gray, and N. Fernández-García, "Architecting Time-Critical Big-Data Systems," IEEE Trans. Big Data, vol. 2, no. 4, pp. 310–324, 2016. doi: 10.1109/TBDATA.2016.2622719.
- [19] W. L. Chang, D. Boyd, and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 6, Big Data Reference Architecture [Version 2]," NIST Special Publication 1500-6r1, 2018. doi: 10.6028/NIST.SP.1500-6r1.
- [20] M. Masmoudi, S. Ben Abdallah Ben Lamine, M. H. Karray, B. Archimede, and H. Baazaoui Zghal, "Semantic Data Integration and Querying: A Survey and Challenges," ACM Comput. Surv., vol. 56, no. 8, Art. 209, pp. 1–35, 2024. doi: 10.1145/3653317.