Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

Modular and Extensible Framework for Real-Time Social Media Analytics: Modeling Functional Requirements

Fairouz Zendaoui*

*Corresponding author, Laboratoire de la Communication dans les Systèmes Informatiques, Ecole Nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. E-mail: f_zendaoui@esi.dz

Walid Khaled Hidouci

Laboratoire de la Communication dans les Systèmes Informatiques, Ecole Nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. E-mail: w_hidouci@esi.dz

Journal of Information Technology Management, 2025, Vol. 17, Issue 3, pp. 197-216
Published by the University of Tehran, College of Management
doi:%20https://doi.org/ 10.22059/jitm.2025.104046
Article Type: Research Paper

Received: February 28, 2025
Received in revised form: April 03, 2025
Accepted: July 16, 2025
Published online: August 23, 2025

Abstract

Social media platforms have become essential sources for studying social, political, and cultural dy-namics. However, current solutions remain fragmented, domain-specific, and often inaccessible to non-experts. This paper introduces an innovative, generic, and extensible conceptual model to bridge this gap. Unlike existing approaches, our framework offers a unified architecture integrating real-time data collection, structured storage, multilingual processing, search, and semantic analysis (in-cluding sentiment analysis and beliefs) within a modular system. It ensures adaptability to the di-verse needs of researchers and professionals. This model stands out through its unified workflow (collection, analysis, and visualization), turnkey interface for non-experts, and extended semantic capa-bilities. We identify critical functional requirements through a comprehensive review of existing tools, highlighting their limitations. We then model a system of independent yet interoperable mod-ules: real-time stream management, filtering, automatic classification (sentiment, topics), and exten-sion mechanisms. While conceptual, this model lays the foundation for practical implementation, illustrated by use cases that show its relevance in research and industry. Designed to meet research-ers' needs, it opens promising avenues for analyzing public opinion and digital behaviors in social media studies.

Keywords: Social Media Data, Real-Time Data Processing, Generic Conceptual Model, Modular Architecture, Semantic Analysis.

Introduction

Social media has become a major source of data for both academic and industrial research (Nugroho & Angela, 2024; Zulfakhairi Mokthar et al., 2025), offering unique insights into digital behaviors (Xue et al., 2023), social dynamics (Joshi, 2023), and public opinion (Rodríguez-Ibánez et al., 2023). The rise of platforms like Twitter has generated a massive volume of real-time data, creating unprecedented opportunities to analyze social, political, economic, and cultural events (Singh et al., 2024). However, this abundance of data comes with technical and methodological challenges, particularly in terms of data collection, storage, management, and large-scale analysis (Camacho et al., 2020). Despite the variety of models developed for the collection and analysis of social media data (Bacha et al., 2023; Khan & AlGhamdi, 2024; Wahid et al., 2022; Zendaoui & Hidouci, 2024), many exhibit significant functional limitations and often remain narrowly focused, tailored to specific use cases. Although efficient in extracting large volumes of data, some tools lack real-time collection capabilities, limiting their usefulness in analyzing instant phenomena such as reactions to breaking news or emerging trends. For example, batch-processing-based models do not allow interaction with continuous data streams, which is essential in contexts such as crisis management or online advertising campaign analysis.

Other models, meanwhile, focus solely on data collection and storage, overlooking advanced analytical features. They do not include mechanisms for performing sophisticated lexical processing, such as semantic analysis or classification by sentiment or belief. This lack of integrated analysis often forces users to rely on third-party tools, leading to fragmented workflows and reduced efficiency. These shortcomings highlight the need for a unified framework capable of simultaneously addressing the requirements of social media data collection, management, and analysis.

This paper proposes a generic and flexible model for the collection and analysis of social media data. The model is based on the identification of functional requirements and the definition of several essential features for the exploitation of social media data, particularly real-time collection, which allows for the rapid analysis of trends as they emerge, especially in critical situations such as political crises or natural disasters. Next, structured data storage ensures data accessibility and organization, thereby facilitating analysis. The model also offers parameterized search and advanced filtering features, enabling the extraction of specific data subsets based on criteria such as geolocation or language. To make social media textual data actionable, the model integrates lexical processing and semantic analysis, which help clean, normalize, and better understand the meaning of messages. Finally, analytical classification of data (sentiment, beliefs, etc.) is included to better capture the opinions and beliefs expressed, which is crucial in fields such as public opinion studies or strategic monitoring. This unified, modular, and extensible model addresses the diverse needs of researchers and professionals while offering a conceptual and methodological framework

suitable for multiple applications, such as social behavior analysis, real-time event monitoring, or the development of data-driven strategies. This work creates opportunities for more efficient exploitation of large-scale social media data and lays the groundwork for future research in this ever-evolving field.

This paper is structured as follows. The State of the Art reviews existing tools and identifies their limitations. The Methodology details the functional requirements, system actors, use cases (including a real-world earthquake response scenario), modular architecture, and technical stack. Next, Applications and Usage Perspectives highlights the model's innovative positioning, potential domains, and future extensions. The Conclusion summarizes key contributions and outlines future research directions.

State of the Art

Collection and Analysis of Social Media Data

The collection and analysis of data from social media has been the subject of numerous research efforts, which can be categorized into several main areas. The first category concerns real-time data extraction. Research in this domain focuses on the use of APIs, web scraping techniques, and processing pipelines to retrieve real-time data from platforms such as Twitter, Facebook, or Instagram. These studies also address challenges related to scalability, particularly managing the massive data streams generated by these platforms, which requires solutions adapted to growing data volumes and high-performance demands (Mussina et al., 2020). Another important category is content processing and analysis (Bhandari et al., 2023). This approach includes studies on metaphor extraction, named entity recognition, hashtag analysis, and the identification of trending topics. These efforts aim to extract relevant information from content shared on social media, thus enhancing the understanding of online conversations, identifying key events, and mapping areas of interest. Textual data analysis in this context is crucial for extracting underlying meanings and trends, and this remains a highly active research area. Specific applications of social media data analysis also constitute a major research focus. Researchers have particularly explored areas such as public opinion analysis (Rouhani & Abedin, 2020), fake news detection (Aïmeur et al., 2023), uncertainty identification (Zendaoui et al., 2022), the study of collective behaviors, and live event monitoring. These applications help understand and predict social phenomena, analyze public reactions to events, and combat the spread of misinformation. These studies emphasize the importance of using social media data for societal and political analysis. Finally, the exploration of linguistic and cultural dimensions represents another key area. These works focus on the peculiarities of multilingual and multicultural data, as well as the implications of these differences for the accuracy of analyses (Lin et al., 2018). Researchers analyze how linguistic and cultural variations influence how users interact with social platforms and how these factors can affect data interpretation. This field of research is particularly relevant in an

increasingly globalized world, where social media users come from diverse cultures and languages, presenting additional challenges for analyzing their behaviors and opinions.

Thus, the state of the art in the collection and analysis of social media data highlights the diversity of tools and platforms used to process large and complex data. Researchers deploy a variety of approaches tailored to the specifics of each platform and analysis objective, which requires the creation of dedicated tools for each context to ensure solid foundations for research. This need for specific models and tools is present at every stage of the data collection, analysis, and interpretation process. It is also essential to emphasize the importance of increased support to assist researchers in these complex tasks by providing frameworks, libraries, and collaborative platforms that facilitate data management, integration of advanced models, and optimization of analysis processes. This support is crucial to accelerate scientific advancements and ensure the reliability of results obtained in a constantly evolving environment. Table 1 presents a typology of tools commonly used for social media data analytics, categorizing them based on their primary functions and target users. These tools vary in terms of complexity, functionality, and intended audience, ranging from open-source platforms suited for detailed social graph analysis and visualization to commercial software-as-a-service (SaaS) tools designed for sentiment analysis and brand monitoring. Developer-oriented tools provide raw data collection through APIs, offering flexibility for custom processing, while aggregation platforms and dashboards allow for multi-network monitoring and real-time tracking. Finally, academic and research tools are tailored for targeted data extraction and visualization, often used for social or political analysis. This table highlights the diversity of tools available and their specific applications in the realm of social media data analytics.

Table 1. Typology of Tools for Social Media Data Analytics

Category	Tools	Main Functions
Open-source tools	Gephi, NodeXL, Netlytic	Social graph analysis and visualization
Commercial (SaaS) tools	Brandwatch, Talkwalker, Sprinklr	Sentiment analysis, brand monitoring, and automated reporting
Developer-oriented tools	Tweepy (Python), Twarc, snscrape	Raw data collection via APIs for custom processing
Aggregation platforms & dashboards	Hootsuite, Buffer, Keyhole	Multi-network monitoring, real-time tracking
Academic / research tools	SocioViz, Crimson Hexagon (formerly), TAGS (Twitter Archiving Google Sheet)	Targeted extraction and visualization for social or political analysis

To effectively compare the various platforms for collecting and analyzing data from social media, it is crucial to establish a well-defined set of evaluation criteria. These criteria serve as the foundation for assessing the capabilities and limitations of each platform, enabling a more thorough and objective comparison. The primary goal is to evaluate the technical strengths, user-friendliness, feature set, and adherence to ethical standards that the platforms offer. By carefully considering these aspects, researchers and developers can choose

the most suitable platform for their needs. The selected evaluation criteria cover key dimensions essential for assessing a platform's effectiveness. Accessibility considers whether the platform is open-source or commercial and how user-friendly it is, balancing flexibility with ease of use. Covered Sources refers to the range of social media and data sources supported, enhancing the breadth of analysis. Features include tools like filtering, visualization, sentiment analysis, and community detection, which help extract meaningful insights. Multilingual Support is critical for cross-cultural research, allowing analysis beyond English-speaking regions. Compliance with API Usage Policies ensures ethical, privacy-respecting, and legal data use. Finally, Customization allows users to tailor metrics, visualizations, and workflows to meet specific research needs, boosting the platform's relevance and efficiency.

Limitations of Current Solutions

Research on collecting and analyzing social media data has advanced in areas like real-time extraction, content analysis, and applications to social and political issues. However, these studies reveal a key gap: the lack of universal, standardized models. This limitation is clear, given the complexity of current tools, which require advanced technical skills. Existing tools often demand programming or data science expertise, limiting accessibility for researchers from diverse disciplines. This restricts data use in fields like social sciences, crisis management, and political and cultural studies, underscoring the need for more accessible models. Examples highlight this issue: during political crises, tracking real-time conversations on Twitter or Facebook is crucial, but current tools require technical expertise to configure filters or sentiment analysis, excluding many social science researchers. Similarly, studying reactions to an earthquake through geolocated tweets is hampered by complex interfaces. Analyzing misinformation, such as tracking fake news, also requires constant algorithm adjustments, making research difficult without programming skills. Thus, the absence of a universal model limits the use of real-time analysis tools and their impact across various fields. To address this, we propose a generic, extensible model for collecting and analyzing social media data. This solution offers researchers a ready-to-use, adaptable tool that requires no advanced programming, expanding access to real-time data analysis. Our approach aims to democratize social media analysis, providing a simple alternative for research on political events, disasters, and social phenomena

Methodology

Identification of Functional Requirements

The identification of functional requirements is a crucial step in designing a generic and extensible model for the real-time collection and analysis of data from social media platforms. This phase aims to clearly define the functionalities needed to ensure the effective use of social information flows across various application contexts. To identify the main trends in

the field and better understand the functional needs related to the collection and analysis of social media data, we conducted an in-depth study of existing approaches, available tools, and the specific requirements of targeted applications. Following this analysis, several major functional requirements and key features were identified for the collection and analysis of data from social networks.

- **Real-Time Data Collection**: The system must collect continuous data from platforms like Twitter, Facebook, Instagram, and TikTok, managing live streams and ad hoc queries, while accounting for each platform's APIs and restrictions.
- **Management of Multilingual and Heterogeneous Data**: It must integrate advanced NLP to analyze multilingual content and consider cultural specificities. Harmonizing formats and structures ensures consistent, reliable analysis.
- **Advanced Data Filtering and Search**: The system should offer advanced filtering by time, language, geolocation, keywords, hashtags, and mentions, using optimized algorithms for targeted analysis.
- **Integration of Lexical Processing**: Lexical processing (noise removal, normalization, correction) is key to improving data quality. Specialized lexicons and knowledge bases enhance semantic analysis, entity recognition, and accuracy.
- **Analytical Data Classification**: AI and machine learning models should classify data by sentiment, topics, influential opinions (disinformation, polarization), and emerging trends, adapted to user contexts.
- **Administration and Configuration**: The system needs intuitive admin tools for settings, roles, data access, and monitoring, supporting dynamic parameter configuration without code changes. Secure authentication, role-based control, and logging ensure governance and compliance.
- **Adaptability and Extensibility**: The modular, flexible architecture must support new features, handle growing data volumes, and comply with evolving regulations, ensuring adaptability for media monitoring, crisis management, or marketing analysis.

By addressing these functional requirements, the model offers a robust, flexible, and scalable solution for leveraging social media data. Thanks to its modular and efficient architecture, it provides researchers and professionals with a powerful tool capable of adapting to the specific requirements of each application domain while anticipating technological and regulatory changes. This approach ensures that the system aligns with the diverse needs of users in the field of social data collection and analysis.

Identification of System Actors

As part of the design of a system for collecting, processing, and analyzing data from social media, identifying the types of actors is essential to tailor functionalities to the specific needs of each user profile. These actors can be categorized according to their role in the system, whether technical, analytical, operational, or strategic. Table 2 below presents a synthetic typology of the actors involved at various levels of the system.

Table 2. Typology of system actors

Type of Actors	Examples
Technical	Data Engineer, Developer, Data Architect, System Administrator
Analysts & Scientists	Data Scientist, Data Analyst, Researcher, Consultant, Linguist
Domain Experts	AI Expert, NLP Engineer, Anthropologist, Business/Industry Expert
Business End Users	Journalist, Community Manager, Marketing Manager, CRM Manager
Decision Makers & Strategists	Intelligence Manager, R&D Manager, Communication Manager, Product Strategist
Operational Staff	Moderator, Customer Experience Manager, Marketing Analyst

Definition of Use Cases

As part of our methodological approach, we defined a set of use cases to explain the interactions between the different system actors and the required functionalities. Based on the identified functional requirements, we established several major use cases, which we grouped into packages. Structuring the use cases into functional categories allows us to ensure comprehensive coverage of the identified needs.

1. Real-Time Data Collection

This category covers continuous or on-demand acquisition of data across platforms like Twitter, Facebook, Instagram, and TikTok. It involves setting up API connectors, managing continuous data streams with error handling, and applying initial filters during collection (such as hashtags or keywords). Researchers and analysts may also need ad hoc data retrieval for specific analyses, while system administrators handle API quota management to avoid exceeding request limits. These tasks typically involve data engineers, analysts, data scientists, and developers, depending on the project's needs (Table 3).

Table 3. Real-Time Collection Use Cases

Use Case	Involved Actors	Typical Scenario
Multiple API Connections	Data Engineer, Social Media Analyst	Configuration of API connectors for Twitter, Facebook, etc., to retrieve normalized data in a unified format.
Continuous Data Streams	Data Scientist, Monitoring Manager	Real-time monitoring of brand mentions with automatic stream interruption handling.
Initial Filtering During Collection	Marketing Analyst, Community Manager	Targeted collection of posts containing a specific hashtag or keyword to reduce noise.
On-Demand Collection	Researcher, Consultant	Extraction of historical data for retrospective analysis without using a continuous stream.
API Quota Management	System Administrator, Developer	Monitoring and dynamically adjusting API requests to avoid rate limit issues.

2. Management of Multilingual and Heterogeneous Data

Social media messages differ widely in language and format, making standardization essential. This includes automatic language detection (often using machine learning) to identify texts in languages like French, Spanish, or Arabic, and multilingual processing that integrates translation tools while preserving nuances and idiomatic expressions. It also involves normalizing data formats from different platforms (e.g., JSON, XML, CSV) and merging sources like Twitter and Facebook into a unified dataset for consistent analysis. Additionally, cultural adaptation techniques help interpret local expressions, emojis, or cultural references, reducing the risk of misinterpretation across regions. These tasks typically involve data scientists, NLP engineers, data engineers, analysts, and experts in cultural or linguistic adaptation, depending on the project (Table 4).

Table 4. Multilingual and Heterogeneous Data Management Use Cases

Use Case	Involved Actors	Typical Scenario
Language Detection	Data Scientist, Linguist	An analyst receives tweets in multiple languages. The system automatically detects each language (e.g., French, Spanish, Arabic) using an ML model, enabling pre-sorting for analysis.
Multilingual Processing	NLP Engineer, Internationalization Manager	Automatic translation of Facebook comments into English for centralized analysis, while preserving linguistic nuances (e.g., idiomatic expressions).
Format Normalization	Data Engineer, Data Architect	Conversion of raw data (Twitter in JSON, Facebook in XML) into a standardized CSV format, with common fields (date, author, content).
Data Source Fusion	Business Analyst, Consultant	Aggregation of Instagram and LinkedIn posts on the same topic into a single dataset, with deduplication and metadata alignment.
Cultural Adaptation	Digital Anthropologist, Global Marketing	Detection of cultural connotations in emojis or local expressions (e.g., 👍 interpreted differently in Asia and the West) to avoid interpretation bias.

3. Advanced Data Filtering and Search

After collection and normalization, data must be easily searchable and usable (Table 5). This involves implementing parameterized queries that let users filter by multiple criteria, like date, keywords, geolocation, language, hashtags, or mentions, such as a journalist searching tweets about the “2024 elections” in France with over 100 shares. Semantic search goes beyond exact terms to capture related meanings, for example, identifying posts about “green energy” that mention “ecological transition” or “photovoltaic solar.” Suggestion engines help by proposing trending keywords or hashtags, supporting tasks like brand monitoring. Users can also configure how results are displayed—such as tables, charts, or word clouds—and sort them by criteria like date or user profile, which helps prioritize analysis, for example, when reviewing Facebook comments.

Table 5. Filtering and Search Use Cases

Use Case	Involved Actors	Typical Scenario
Parameterized Queries	Data Analyst, Investigative Journalist	A journalist searches for all tweets containing the keywords "2024 elections" posted in France between January and March 2024, with at least 100 shares. The package allows combining these filters (date, geolocation, engagement) into a single query.
Semantic Search	Social Sciences Researcher, R&D Manager	A researcher wants to analyze discussions on "green energy" without limiting to the exact term. The package identifies posts referring to the subject with synonyms or contextual formulations (e.g., "ecological transition," "photovoltaic solar").
Suggestion Engines	Community Manager, Marketing Strategist	A Community Manager wants to monitor trends around their brand. The package suggests emerging hashtags or keywords (e.g., #NewProduct2024) based on recent influencer activity.
Result Format Configuration	Data Visualist, Consultant	For a client report, a consultant exports the results as a clickable word cloud (highlighting frequent terms) and an Excel table with full metadata (author, date, and source).
Sorting Results	CRM Manager, Platform Moderator	A moderator sorts reported Facebook comments by chronological order (from the most recent) and by profile (verified users first) to prioritize their analysis.

4. Integration of Lexical Processing

Effective data analysis requires thorough text preprocessing (Table 6). This includes cleaning data by removing noise such as stopwords, emojis, and typos (e.g., converting “keuf” to “neuf” or 🐌 to “slowness”), correcting spelling, and standardizing abbreviations (“mdr” → “laughing out loud”). Normalization and lemmatization help reduce linguistic variability, grouping forms like “courais,” “courir,” and “course” under the root “cour.” Semantic analysis uncovers the meaning of texts and identifies relationships between concepts (like linking “iPhone” with “battery problem”). Named entity recognition extracts key entities such as people, companies, or places, useful for tasks like mapping political networks. Finally, lexical enrichment adds depth to analysis by integrating specialized lexicons or external knowledge bases, for example, using medical terms when studying patient forums.

Table 6. Lexical Processing Use Cases

Use Case	Involved Actors	Typical Scenario
Text Data Preprocessing	Data Scientist, Linguist	An analyst cleans tweets containing typos ("keuf" → "neuf"), emojis (🐌 → "slowness"), and abbreviations ("mdr" → "laughing out loud") to standardize the corpus before analysis.
Normalization and Lemmatization	NLP Engineer, Researcher	Reduction of linguistic variations ("courais", "courir", "course" → common root "cour") to improve thematic analysis of a health forum.
Semantic Analysis	Monitoring Manager, Consultant	Automatic identification of dominant topics and relationships between concepts ("iPhone" frequently associated with "battery" and "problem") in customer reviews.
Named Entity Recognition	Journalist, Geopolitical Analyst	Automatic extraction of political figures, companies, and locations from news articles to map relationships.
Lexical Enrichment	Sectoral Marketing, Domain Expert	Adding specialized terms (e.g., medical vocabulary for analyzing patient forums) via integration of external knowledge bases.

5. Analytical Data Classification

This phase aims to extract meaningful insights to support analysis and decision-making. Sentiment classification uses machine learning models to label data as positive, negative, or neutral, helping, for example, a brand analyze thousands of customer reviews for improvement areas (Table 7). Thematic classification groups messages by topics like politics, economy, or health, which is useful for researchers studying public concerns. Belief classification goes further by capturing nuances such as doubt, conviction, or ideological stance. Systems also detect influential opinions, flagging accounts that amplify viral falsehoods or biased content. Emerging trends, such as the rise of hashtags (e.g., #Eco-Fashion), are identified early through pattern analysis. Users can customize classification criteria, such as refining models to detect e-commerce delivery complaints. Other critical capabilities include the detection of risky speech, which automatically alerts moderators to hate or violent remarks, and event detection, which monitors social streams in real-time to capture weak signals of major events, such as sudden spikes in tweets mentioning "fire" in a specific location.

Table 7. Analytical Classification Use Cases

Use Case	Involved Actors	Typical Scenario
Sentiment Classification	Customer Experience Manager, Market Analyst	A brand automatically analyzes 10,000 customer reviews on Twitter, classified as positive/negative/neutral, to identify product improvement areas.
Thematic Classification	Journalist, Social Science Researcher	Automatic categorization of 50,000 Reddit posts by themes (economy, environment, health) for a study on post-crisis public concerns.
Detection of Influential Opinions	Communications Manager, Geopolitical Analyst	Identification of Twitter accounts amplifying a viral falsehood via analysis of sharing networks and biased linguistic patterns.
Detection of Emerging Trends	Marketing Strategist, Data Scientist	Early detection of the hashtag #EcoFashion in Instagram discussions before it goes mainstream, through analysis of abnormal growth curves.
Customization of Criteria	Industry Consultant, CRM Manager	Adaptation of a generic sentiment model to specifically detect complaints about delivery delays in e-commerce.
Detection of Risky Speech	Content Moderator, HR Manager	Automatic alerts for messages containing racial slurs or calls to violence in a company's YouTube comments.
Event Detection	Security Analyst, Crisis Manager	Real-time identification of an abnormal spike in geolocated tweets containing "fire" + "Lyon" to alert emergency services.

6. Administration and Configuration

This module ensures efficient management of users and system settings (Table 8). It includes user profile and role management, enabling administrators to create, modify, or delete accounts and assign precise access rights, such as limiting a junior analyst's access to recent data without export permissions. Search and analysis criteria can be customized—for instance, a consultant may set permanent filters to monitor topics like "energy transition" in French-language posts. Algorithm updates are supported, allowing technical teams to improve

models (e.g., adding source verification to a disinformation detector) without disrupting service. System configuration management lets IT teams adjust global parameters like processing frequency or detection thresholds, adapting the system to organizational needs. Finally, the platform allows users to save and reuse personalized configurations, such as a predefined crisis monitoring setup, to optimize workflows and ensure consistency across teams.

Table 8. Administration and Configuration Use Cases

Use Case	Involved Actors	Typical Scenario
User Profiles and Role Management	System Administrator, Security Manager	An administrator creates a new "Junior Analyst" profile with limited access to data from the past 6 months and no export rights, in accordance with company policy.
Customization of Search and Analysis Criteria	Sector Analyst, Consultant	An energy consultant sets up a permanent filter to simultaneously track the terms "energy transition," "renewable energies," and "sobriety" in French-language posts.
Algorithm Updates	Data Scientist, ML Engineer	The technical team improves the disinformation detection model by integrating a new source verification module, without service interruption.
System Configuration Management	IT Manager, Solution Architect	Adjustment of processing parameters to reduce analysis frequency from 15 to 60 minutes during peak social media activity periods.
Configuration Saving	Senior Analyst, Intelligence Manager	An analyst saves their optimal crisis monitoring setup (predefined keywords, alert thresholds) and shares it with their team to ensure a uniform response.

7. Adaptability and Extensibility Management

To ensure the system's long-term sustainability and adaptability, it incorporates several extensibility mechanisms (Table 9). The modular architecture design facilitates easy updates and the addition of new features without overhauling existing components, such as integrating a new image analysis module via standardized APIs. Horizontal scalability is built in, allowing resources like servers or databases to scale up automatically—this is especially useful during events like marketing campaigns, which generate surges in data. The system can also incorporate new data sources, such as emerging social media platforms like Mastodon, in a short timeframe (e.g., two days following user migration from Twitter). New user profiles can be defined, such as the creation of an "Auditor" role with limited access to data for compliance purposes. The architecture supports the rapid addition of new platforms, like Threads, through prototyping. Additionally, regulatory compliance is ensured by automatically updating data retention policies to align with new legal requirements. Lastly, extensibility testing, such as load testing to simulate a high volume of simultaneous requests, ensures the system can handle peak demands without performance degradation.

Table 9. Adaptability and Extensibility Management Use Cases

Use Case	Involved Actors	Typical Scenario
Modular Configuration	Software Architect, DevOps	Integration of a new image analysis module without refactoring existing code, using standardized API interfaces.
Horizontal Scalability	Cloud Engineer, Infrastructure Manager	Automatic scaling from 5 to 20 processing nodes during the launch of a marketing campaign that generates a data surge.
Integration of New Sources	Integration Developer, Product Owner	Adding Mastodon support to the system in 2 days following user migration from Twitter.
User Profile Management	Security Officer, System Admin	Creation of an "Auditor" role with limited access to personal data for compliance purposes.
Support for Emerging Platforms	Technology Watch Team, R&D Team	Rapid prototyping of a connector for Threads (Meta), three weeks after its official launch.
Regulatory Compliance	DPO (Data Protection Officer), Legal Advisor	Automatic update of data retention rules to comply with new European legislation.
Extensibility Testing	QA Lead, Performance Engineer	Simulation of 1 million simultaneous requests to validate system stability before Black Friday.

Modular Architecture of the Framework

The modular architecture we proposed, as illustrated in Figure 1, is structured around autonomous functional blocks interconnected via a central orchestration layer. This design enables a clear separation of concerns while facilitating coordination between components. Each module — including Data Collection, Linguistic Processing, Search, Classification, and Administration— is assigned specific responsibilities and exposes standardized interfaces. Such an organization enhances system maintainability and supports seamless evolution, allowing modules to be updated or replaced independently. At the core of this architecture, the API gateway serves as a single entry point for all external requests, whether originating from users or third-party applications. It also handles key cross-cutting concerns such as access authentication, request routing, and interaction monitoring. Data is stored and managed within a unified Data Lake, which constitutes the system's persistent layer. Initially recorded in raw form, data is progressively enriched through a series of processing stages, including lexical cleaning, semantic analysis, and classification. This pipeline approach ensures traceability and facilitates data reuse throughout the processing chain. Communication between modules follows a dual mechanism: asynchronous flows are employed for real-time data processing, ensuring system responsiveness, while synchronous calls are used for user interactions, providing reliable and immediate feedback. To ensure global system coherence, a centralized configuration management system is employed. It enables the consolidation of operational parameters, monitors the status of all modules, ensures component compatibility, and supports coordinated updates.

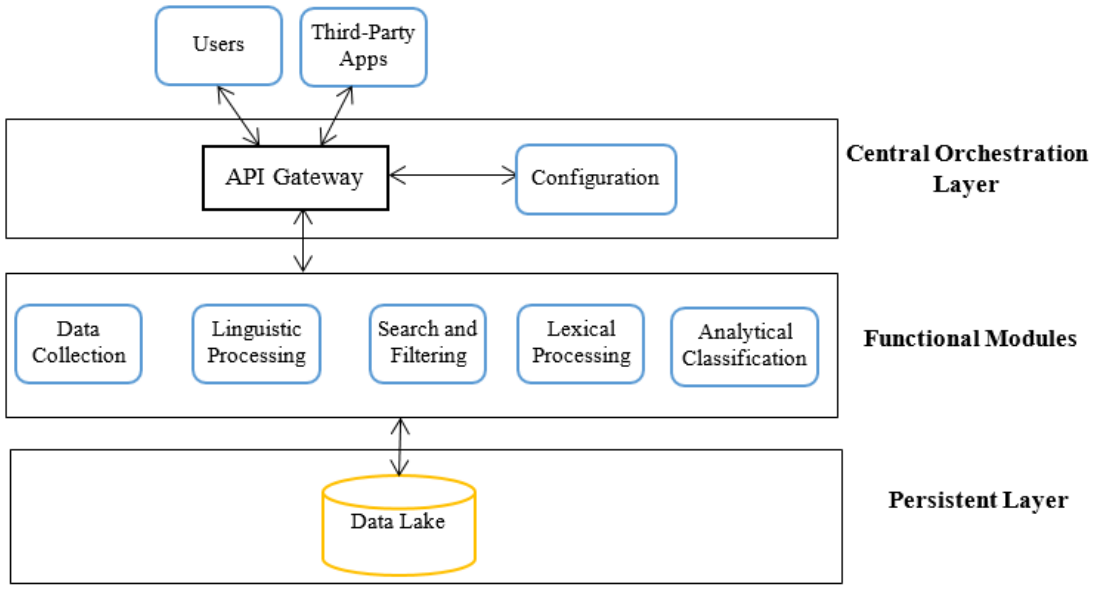

Figure 1. Modular Architecture of the Framework

This modular architecture provides strong flexibility and scalability. New sources or functionalities — such as connectors for emerging social media platforms — can be integrated without overhauling the entire system. Moreover, module-based horizontal scalability ensures optimal performance even under high load conditions

Technical Stack of the Framework Prototype

The prototype to be developed is based on a modern and modular technology stack, designed to meet the requirements of flexibility, advanced analysis, and accessibility identified in the theoretical framework (Table 10). At its core, an API Gateway (Kong or NGINX) manages authentication, routing, and monitoring, while asynchronous communication between modules is handled by a message broker like Apache Kafka or RabbitMQ. Data is stored in a Data Lake (Amazon S3 or Azure Blob Storage) and processed via a data pipeline using Apache Spark or Flink, enabling efficient real-time enrichment. Functional modules include real-time data collection (Tweepy, Twitter API), linguistic processing (spaCy, NLTK), search (Elasticsearch, Apache Solr), and classification (TensorFlow, PyTorch, BERT). Configuration management is centralized using HashiCorp Consul or Spring Cloud Config, while system performance and health are monitored with Prometheus/Grafana and the ELK Stack. The user interface is built with ReactJS or Vue.js, integrating push notifications via Twilio and Firebase for real-time alerts. Security is ensured through OAuth 2.0 and JWT, providing secure access control. This modular architecture allows for easy integration of new modules and data sources, ensuring scalability and robustness for large-scale data processing and system evolution.

Table 10. Technological Stack of the Prototype

Component	Technology	Responsibilities	Features
API Gateway	Kong API Gateway, NGINX	Manages requests, authentication, routing, and monitoring	Security (JWT, OAuth), request routing, logging
Orchestration Layer	Apache Kafka, RabbitMQ	Handles asynchronous communication between modules	High-throughput, low-latency message handling
Data Lake	Amazon S3, Azure Blob Storage	Stores raw, unstructured data	Scalable storage, high availability
Data Processing Pipeline	Apache Spark, Apache Flink	Real-time data enrichment (lexical cleaning, semantic analysis)	Distributed, scalable processing
Data Collection	Tweepy, Twitter API (V2)	Collects social media data based on keywords, hashtags	Real-time streaming, API integration
Linguistic Processing	spaCy, NLTK	Text processing (tokenization, entity recognition)	High-speed NLP, multi-language support
Search	Elasticsearch, Apache Solr	Fast search across collected data	Full-text search, scalable querying
Classification	TensorFlow, PyTorch, Scikit-learn	Text classification (belief analysis, sentiment, etc.)	Hybrid models for belief and sentiment classification
Configuration Management	HashiCorp Consul, Spring Cloud Config	Manages configuration, module status, and version control	Dynamic updates, rollback configurations
Monitoring & Logging	Prometheus, Grafana, ELK Stack	Monitors health, resource usage, logging, and alerting	Real-time metrics, logging, and visualization
User Interface (UI)	ReactJS, Vue.js	Displays processed data and handles user interaction	Dynamic real-time data visualization
Mobile & SMS Notifications	Twilio API, Firebase Cloud Messaging (FCM)	Sends real-time updates to users via mobile or SMS	Cross-platform notifications
Authentication & Security	OAuth 2.0, JWT	Secures access to the system, manages user authentication	Token-based authentication, fine-grained access control

Real-World Use Case: Earthquake Emergency Response

To concretely illustrate the applicability of this architecture and technology stack, we consider the scenario of a natural disaster, such as an earthquake, where real-time data collection, parameterized search, and intelligent classification play a crucial role in accelerating emergency response. From the very first tremors, the system can automatically begin capturing tweets containing specific keywords such as “earthquake,” “rescue,” or “injured.” These are filtered to prioritize trusted sources—local media, official agencies, and verified eyewitness accounts. The incoming data undergoes multiple layers of analysis to ensure a comprehensive understanding and accurate categorization. First, geolocation extraction identifies and converts textual mentions of locations (e.g., “school on Dubois Street”) into precise GPS coordinates. Next, categorization sorts tweets by emergency type, such as medical assistance, logistical needs, or infrastructure damage, providing targeted insights. Finally, disinformation detection assesses the credibility of messages by evaluating source reliability and contextual consistency, ensuring the integrity of the information being analyzed. Once processed, the enriched data are aggregated to produce an interactive crisis

map, displaying key zones of alert, nearby infrastructure (e.g., hospitals, shelters), and dynamic relief operations. This visual interface is continuously updated and automatically shared with field teams via mobile apps or SMS, enabling real-time coordination and optimized prioritization of interventions. This agile and modular analysis chain significantly accelerates decision-making and humanitarian response during the critical first hours following an earthquake. To operationalize this workflow, we formalized it as the following pseudocode, which summarizes the main algorithm executed by the system during an earthquake emergency event. This pseudocode summarizes the system's coordinated workflow for earthquake emergency response, demonstrating the integration of data collection, analysis, and alerting components in real time.

Algorithm 1. Earthquake Emergency Response

Begin

1. Start data stream from social media API (keywords: earthquake, rescue, injured)
2. For each incoming message:
   a. Verify source credibility
   b. Perform geolocation extraction → GPS coordinates
   c. Categorize message → emergency type (medical, logistical, infrastructure)
   d. Detect potential disinformation
   e. Store enriched message in Data Lake
3. Update crisis map with aggregated data: Display alerts, infrastructure, relief operations
4. Notify field teams via mobile/SMS: Send alerts with location and priority
5. Monitor system performance and adjust parameters dynamically

End

Applications and Usage Perspectives

Innovative Positioning of the Model

The distinctive value of this model lies in its threefold contribution: unifying, democratic, and scalable. While current tools (Table 11) require a complex technical chain (APIs → Python → PowerBI), our framework integrates these steps into a single, no-code, customizable dashboard, thereby reducing analysis time for non-experts. Unlike Brandwatch (which is limited to basic sentiment analysis), we incorporate belief analysis through hybrid models (lexicons + transformers), which is crucial for studying social polarization or disinformation campaigns. Finally, its modularity surpasses NodeXL by enabling the addition of new

connectors (e.g., Threads) within 48 hours, thanks to an API-first architecture—a decisive asset in the ever-changing social media ecosystem. We compared the analytical capabilities of Gephi (Bastian et al., 2009; Gephi.org, 2023), NodeXL (Smith et al., 2014; Social Media Research Foundation, 2024), Brandwatch (Mariani et al., 2022; Brandwatch by Cision, 2024), and our model. Unlike Gephi and NodeXL, which require programming or Excel skills, our model relies on a declarative no-code approach, making it more accessible to social science researchers. It also stands out with its low-latency, real-time analysis, crucial for crisis response, and its integration of belief analysis using lexicons and machine learning models, a feature missing in the other tools. Furthermore, it offers multilingual support, enabling cross-cultural studies, and features modular connectors via an open API, ensuring strong adaptability to various platforms.

Table 11. Benchmark of Analytical Capabilities

Criterion	Gephi (Bastian et al., 2009)	NodeXL (Smith et al., 2014)	Brandwatch (Mariani et al., 2022)	Our Model	Key Advantage
No-code interface	✗ (Java/Python)	✗ (Excel)	✓ (Limited)	✓ (Declarative)	Accessibility for social science researchers
Real-time analysis	✗	✗	✓	✓ (Low-latency)	Crisis responsiveness
Belief analysis	✗	✗	✗	✓ (Lexicons + ML)	Disinformation detection
Multilingual support	✗	✗	✓ (Paid)	✓ (multiple languages)	Cross-cultural studies
Modular connectors	✗	✗	✗ (Locked)	✓ (Open API)	Platform adaptability

This benchmarking reveals three gaps addressed by our approach:

- **Democratization**: Unlike Brandwatch (reserved for corporate use) and Gephi/NodeXL (which require technical skills), our model makes advanced analysis accessible through pre-configured templates (e.g., "Humanitarian Crisis Detection").
- **Analytical Depth**: The integration of belief analysis (absent even in Brandwatch) enables the study of complex phenomena such as radicalization or misinformation.
- **Agility**: Its modularity reduces adaptation time to new social networks (e.g., integration of Mastodon in 2 days vs. 6 months for SaaS solutions).

These advancements position our model as a particularly suitable solution for contexts requiring rapid execution, adaptive flexibility, and technical accessibility. University laboratories will find in it a tool meeting budget constraints while enabling complex multilingual analyses essential for interdisciplinary research. NGOs will benefit from its rapid deployment in crisis zones, even with limited connectivity, allowing for informed, real-time

decisions. Media organizations can use it for instant political discourse verification and misinformation detection, reinforcing their mission of delivering verified information. Thanks to its unified architecture, our solution meets academic requirements (reproducibility), operational needs (responsiveness), and ethical standards (transparent analysis), thus covering the full spectrum of needs in social media research and analysis.

Application Domains

Each functional requirement of the model is linked to concrete use cases across various fields:

- **Strategic Monitoring**: Tracking trends and early detection of weak signals within a specific domain (e.g., monitoring discussions around new regulations).
- **Crisis Management**: Rapid identification of emergencies in real-time, such as natural disasters or health crises.
- **Market Analysis**: Leveraging consumer opinions and interactions on social media to assess the perception of a product or service.
- **Content Moderation**: Automatic detection of hate speech or misinformation to improve platform regulation.
- **Study of Digital Behavior and Public Opinion**: In-depth analysis of social dynamics, discursive evolutions, and user interactions.

This approach ensures that the proposed model is both robust and flexible in responding to the various applications of social media data.

Possible Extensions of the Model

The following points outline possible extensions of the proposed model, highlighting avenues for enhancing its analytical capabilities and broadening its applicability.

- **Thematic Modeling and Knowledge Discovery**: Thematic modeling techniques facilitate the identification of emerging themes or topics within online discussions. This allows for in-depth analysis of textual data corpora, enabling researchers to uncover hidden patterns and contextual trends. These tools are particularly useful for studying the evolution of public debates, crisis management, or exploring new research areas. Additionally, they support knowledge extraction processes, where relationships and insights are automatically generated from large-scale collected data, contributing to a better understanding of online social phenomena.
- **Truth Discovery**: This approach enables the processing of information from diverse and potentially contradictory sources by integrating mechanisms to assess the veracity of data.

It enhances the quality of analysis by identifying the most reliable facts and eliminating biases, thus providing more robust results for researchers and practitioners.

- **Adaptability to Other Social Media Platforms**: The proposed generic model could be adapted to include data from other online platforms such as Instagram, LinkedIn, or TikTok. These extensions would require adding new features to handle specific data formats and meet the unique characteristics of each platform.
- **Integration of Predictive Analysis**: An interesting extension would be the integration of predictive analysis modules, using machine learning to anticipate social, economic, or political trends based on data collected in real-time.

Conclusion

This paper presents an innovative, modular, and extensible conceptual model for the real-time collection and analysis of social media data. Designed to address the evolving needs of researchers and professionals, the model integrates previously fragmented functionalities, such as continuous data collection, multilingual processing, and advanced lexical and semantic analysis, including sentiment and belief analysis, all within a flexible and declarative architecture. What sets this model apart is its user-friendly interface, extensibility, and agile design, making it accessible to non-experts while remaining adaptable to the fast-paced digital landscape. Its generic structure broadens the accessibility of social data analytics, enabling diverse disciplines to leverage these tools.

Future work will focus on implementing and deploying a functional prototype, evaluating its usability with non-experts, and refining the system based on empirical feedback. Key research directions include integrating static system analysis for formal verification of performance, reliability, and scalability; optimizing preprocessing algorithms for complex multilingual data; incorporating credibility assessment mechanisms to combat misinformation; and adopting the latest advancements in natural language processing, such as transformers and large language models (LLMs), to enhance semantic and thematic analysis.

From an academic standpoint, this model lays a solid foundation for interdisciplinary research in computational social sciences. In industry, it offers a promising approach to strategic monitoring, critical event detection, and trend analysis. With its versatility, the model can be adapted to various fields, including political studies, humanitarian response, and consumer behavior analysis. Although conceptual, this work paves the way for future empirical validation, including testing for scalability, usability, and robustness before large-scale deployment. Ultimately, this model represents a key step toward the development of intelligent, ethical, and accessible tools for understanding and managing online social dynamics.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining, 13*(1), 30.

Bacha, J., Ullah, F., Khan, J., Sardar, A. W., & Lee, S. (2023). A deep learning-based framework for offensive text detection in unstructured data for heterogeneous social media. *IEEE Access, 11*, 124484-124498.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media, 3*(1), 361-362. https://doi.org/10.1609/icwsm.v3i1.13937.

Bhandari, N., Navalakhe, R. and Prajapati, G. (2023). Social Media Toxic Content Filtering System using SOIR Model. *Journal of Information Technology Management, 15*(Special Issue: Digital Twin Enabled Neural Networks Architecture Management for Sustainable Computing), 78-94. doi: 10.22059/jitm.2023.91569.

Brandwatch by Cision. (2024). Social Media Analytics & Consumer Intelligence. https://www.brandwatch.com/.

Camacho, D., Panizo-LLedot, A., Bello-Orgaz, G., Gonzalez-Pardo, A., & Cambria, E. (2020). The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion, 63*, 88-120.

Gephi.org. (2023). Gephi – The Open Graph Viz Platform (Version 0.10.1). https://gephi.org/.

Joshi, P. (2023). Exploring the Social Dynamics of Health Information Sharing on Social Media: The Influence of Beliefs and Source Information.

Khan, M. A., & AlGhamdi, M. (2024). A customized deep learning-based framework for classification and analysis of social media posts to enhance the Hajj and Umrah services. *Expert Systems with Applications, 238*, 122204.

Lin, B. Y., Xu, F. F., Zhu, K., & Hwang, S. W. (2018, July). Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 709-719).

Mariani, M. M., Machado, I., & Nambisan, S. (2022). Types of innovation and artificial intelligence: A systematic quantitative literature review and research agenda. *Journal of Business Research, 155*(Part A), 113364.

Mussina, A. B., Aubakirov, S. S., & Trigo, P. (2020, November). An Architecture for Real-Time Massive Data Extraction from Social Media. In *International Conference on Mathematical Modeling and Supercomputer Technologies* (pp. 138-145). Cham: Springer International Publishing.

Nugroho, D., & Angela, P. (2024). The impact of social media analytics on SME strategic decision making. *IAIC Transactions on Sustainable Digital Innovation (ITSDI), 5*(2), 169-178.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications, 223*, 119862.

Rouhani, S., & Abedin, E. (2020). Crypto-currencies narrated on tweets: a sentiment analysis approach. *International Journal of Ethics and Systems, 36*(1), 58-72.

Singh, J., Pandey, D., & Singh, A. K. (2024). Event detection from real-time twitter streaming data using community detection algorithm. *Multimedia Tools and Applications, 83*(8), 23437-23464.

Smith, M., Rainie, L., Shneiderman, B., & Himelboim, I. (2014). Mapping Twitter topic networks: From polarized crowds to community clusters. Pew Research Center.

Social Media Research Foundation. (2024). NodeXL Pro: Network Overview, Discovery and Exploration for Excel. https://www.smrfoundation.org/nodexl/.

Wahid, J. A., Shi, L., Gao, Y., Yang, B., Wei, L., Tao, Y., ... & Yagoub, I. (2022). Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response. *Expert Systems with Applications, 195*, 116562.

Xue, Z., Li, Q., & Zeng, X. (2023). Social media user behavior analysis applied to the fashion and apparel industry in the big data era. *Journal of Retailing and Consumer Services, 72*, 103299.

Zendaoui, F., Hidouci, W. K., & Rouhani, S. (2022). Uncertainty identification in microblogs. *Journal of Optimization in Industrial Engineering, 15*(1), 301-309.

Zendaoui, F., Hidouci, W.K. (2024). Considering Uncertainty Expression in Sentiment Analysis and Tweet Classification. In: Shaikh, A., Alghamdi, A., Tan, Q., El Emary, I.M.M. (eds) Advances in Emerging Information and Communication Technology. ICIEICT 2023. *Signals and Communication Technology*. Springer, Cham. https://doi.org/10.1007/978-3-031-53237-5_17.

Zulfakhairi Mokthar, M., Wan Abu Bakar, W. H. R., Shabri, S. and Radzuan, F. A. (2025). The Influence of Social Media on Public Health Protection against the COVID-19 Pandemic through Public Health Awareness and Changes in Behavior: An Integrated Model. *Journal of Information Technology Management, 17*(1), 86-98. doi: 10.22059/jitm.2025.99925.

Bibliographic information of this paper for citing:

Zendaoui, Fairouz & Hidouci, Walid Khaled (2025). A Modular and Extensible Framework for Real-Time Social Media Analytics: Functional Requirements Modeling. *Journal of Information Technology Management,* 17 (3), 197-216. https://doi.org/ 10.22059/jitm.2025.104046