International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Harnessing Big Data and NLP for Real-Time Market Sentiment Analysis Across Global News and Social Media

Rashi Nimesh Kumar Dhenia

Independent Researcher
Email: *dhenairashi[at]gmail.com*

Abstract: ***This research explores a scalable pipeline that utilizes Natural Language Processing (NLP) and Big Data technologies for real-time market sentiment analysis across global news and social media platforms. The exponential growth of unstructured content, ranging from financial headlines to social media opinions, presents an opportunity for businesses to extract actionable intelligence. We propose a hybrid framework that integrates web scraping, text preprocessing, sentiment scoring (using VADER, TextBlob, and BERT), topic modeling (LDA, BERTopic), and real time dashboards via Apache Kafka, Spark, and Power BI. This system supports businesses and financial analysts with continuous, contextual, and high-resolution sentiment streams. The solution's scalability, multilingual support, and high accuracy demonstrate its readiness for real world deployment.***

Keywords: Big Data, Data Analysis, Natural Language Processing, Sentiment Analysis, Real-Time Analytics, Market Intelligence, Topic Modeling, Stream Processing, Power BI

1. Introduction

Historically, financial analysis has relied heavily on structured datasets, including balance sheets, earnings reports, financial ratios, and stock price movements, to assess company performance and predict market trends. These quantitative indicators, while foundational, provide a retrospective view and often lag behind real-world events and investor behavior.

However, the digital age has radically transformed how information is created, distributed, and consumed. The rise of global news portals, blogs, and social media platforms like Twitter and Reddit has generated vast amounts of unstructured textual data, offering real-time windows into public opinion and sentiment. Notably, Bollen et al. [1] demonstrated that aggregated Twitter mood could anticipate movements in the Dow Jones Industrial Average, while Tetlock [7] showed that negative tone in financial news correlated strongly with market downturns. These studies confirm that unstructured discourse often precedes or accompanies market movements, making it a critical, yet underutilized, asset for predictive financial modeling.

Public emotion, social momentum, and media narratives now influence trading behavior at unprecedented speeds. A single tweet from a market influencer, a breaking news headline, or a viral social trend can lead to immediate fluctuations in asset valuations and volatility. Traditional models, built primarily on historical market behavior, struggle to account for these rapidly evolving, sentiment-driven dynamics. As a result, there is a growing demand for systems capable of capturing, interpreting, and reacting to real-time sentiment signals [11][15].

The primary motivation behind this research is to harness the capabilities of modern Natural Language Processing (NLP) and Big Data infrastructures to build a real-time sentiment analysis pipeline that is robust, scalable, and adaptable. By leveraging advanced techniques such as transformer-based language models (e.g., BERT [2], FinBERT [9]), and distributed data processing frameworks like Apache Spark [3] and Hadoop [10], we aim to extract actionable intelligence from millions of global news headlines and social media posts. Foundational technologies such as Word2Vec [4] also support semantic modeling of financial text.

This pipeline is designed to meet several critical criteria: Scalability, to handle high-velocity data streams in production environments. Accuracy, to ensure reliable sentiment classification, even in nuanced or domain-specific contexts, Interpretability, to support transparency and trust in business decision making and Multilingual and cross-platform adaptability, to ensure relevance across global markets [14]

In doing so, this research bridges the gap between raw, unstructured textual data and real-time, data-driven financial insight, equipping investors, analysts, and policymakers with tools for more agile and informed decision-making in a volatile and information-saturated market landscape.

2. Literature Survey

2.1 Foundational Work on Sentiment and Market Behavior

Bollen et al. (2011) were among the first to empirically demonstrate that aggregated public mood, especially from Twitter, could predict stock market movements [1]. Their work validated the hypothesis that emotional tone in textual content

influences economic behavior. Similarly, Tetlock (2007) found that negative sentiment in financial news headlines correlated strongly with subsequent downturns in stock prices [7]. These findings have since inspired a generation of sentiment-based financial models [6][9][15].

2.2 Advances in NLP Techniques

The field of NLP has seen dramatic improvements with the introduction of transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. [2]. These models allow for better contextual understanding of sentences, capturing nuanced sentiment with higher accuracy than traditional lexicon-based methods. BERTopic [6] further enhanced topic extraction by leveraging class-based TF-IDF and embeddings. FinBERT, a domain-specific adaptation, provides improved sentiment classification in financial texts [9]. Similarly, Word2Vec [4] laid foundational work in learning dense vector representations crucial to downstream tasks in NLP pipelines.

2.3 Big Data Infrastructure for Real-Time Analytics

Real-time sentiment analysis demands scalable infrastructure. Technologies like Apache Kafka and Apache Spark [3] have become industry standards for handling high-velocity data streams. These tools enable the continuous ingestion, processing, and querying of large datasets with minimal latency, making them essential for time-sensitive market applications. Hadoop [10] complements this architecture for distributed data storage and batch processing, providing the backbone for many large-scale data systems.

3. Problem Definition

3.1 Industry Challenges in Market Sentiment Extraction

Despite the availability of advanced tools, organizations still struggle to implement real-time sentiment analysis due to technical and strategic gaps. Data is often siloed, making it difficult to aggregate across platforms. Lexicon-based models like VADER [5] and TextBlob [8] fail to capture nuanced language like sarcasm or regional dialects. Moreover, many sentiment tools are not designed to process streaming data efficiently or integrate seamlessly with business intelligence dashboards.

3.2 Objectives of the Proposed Framework

This research seeks to address these issues by developing a modular, scalable architecture that combines web scraping, natural language processing, sentiment modeling, and visualization. The goal is to provide a unified solution capable of: Capturing sentiment in real-time from multiple platforms, Using both rule-based and AI-driven sentiment classifiers [2], [5], [9]. Modeling thematic trends using topic modeling [6], [13], Presenting insights through interactive dashboards for real-time decision-making.

4. Methodology

4.1 Data Collection

News data was gathered using the GNews API, which provides aggregated headlines from global financial outlets. The headlines were retrieved with timestamps and source metadata to preserve temporal and contextual relevance. Using the Twitter API and web scraping techniques [12], tweets containing finance-related hashtags and keywords were collected. Special care was taken to handle rate limits and duplicate entries using tweet IDs and language filters.

4.2 Data Preprocessing

Text data was converted to lowercase, punctuation and HTML tags were stripped, and stop words were removed using the Natural Language Toolkit (NLTK) [7]. Lemmatization was performed using spaCy [11] to standardize word forms. Non-English text was filtered out for this version of the pipeline using langdetect [17]. In future versions, multilingual support will be enabled using models such as XLM-R [18].

4.3 Sentiment Analysis Models

VADER [5] and TextBlob [8] were employed for quick sentiment scoring of short-form texts such as tweets. These models return polarity scores and sentiment labels (positive, negative, neutral). FinBERT [9], a fine-tuned BERT model for financial sentiment, was used to classify longer and more complex sentences. Its contextual embeddings allow for better interpretation of ambiguous or jargon-heavy language.

4.4 Topic Modeling

LDA [13] was used to identify clusters of co-occurring terms, revealing latent themes in the text. This unsupervised approach is effective for traditional corpora but may lose context in short-form content. BERTopic [6] combines transformers and class-based TF-IDF to produce more coherent and contextual topics, particularly useful for social media data. Topics were labeled and visualized using word clouds and frequency distributions.

4.5 Big Data Pipeline

Kafka [3] facilitated real-time ingestion of news and tweet streams into the system. Messages were serialized using Avro [19] for compression and schema management. Structured Streaming was used to process and transform the data into structured formats. Spark [3] also handled joins between sentiment scores, metadata, and time-based aggregations.

4.6 Dashboard and Visualization

Interactive dashboards were created using Power BI [16], enabling stakeholders to view real-time trends, filter by sentiment category, region, or topic, and track changes over customizable time windows. Key visuals include line charts for sentiment over time, heat maps by geography, and topic clouds. The dashboards are refreshed at 5-minute intervals with API integration.

5. System Architecture

5.1 Overall Architecture Design

The proposed system integrates modular components for ingestion, preprocessing, analysis, and visualization. It ensures low-latency throughput by separating streaming and batch jobs, while maintaining a shared schema registry for standardization. Kafka [3] acts as the backbone for ingestion, Spark [3] handles real-time data transformation, and the Power BI service [16] presents actionable visuals to business users.

5.2 Cloud Infrastructure Deployment

The architecture is deployed on AWS [14] with auto-scaling EC2 instances for NLP tasks, S3 for data lake storage, and Amazon EMR for distributed Spark jobs. Kafka is hosted on Amazon MSK to support event streaming with automatic fault tolerance and monitoring. All components are containerized using Docker and orchestrated via Kubernetes [15].

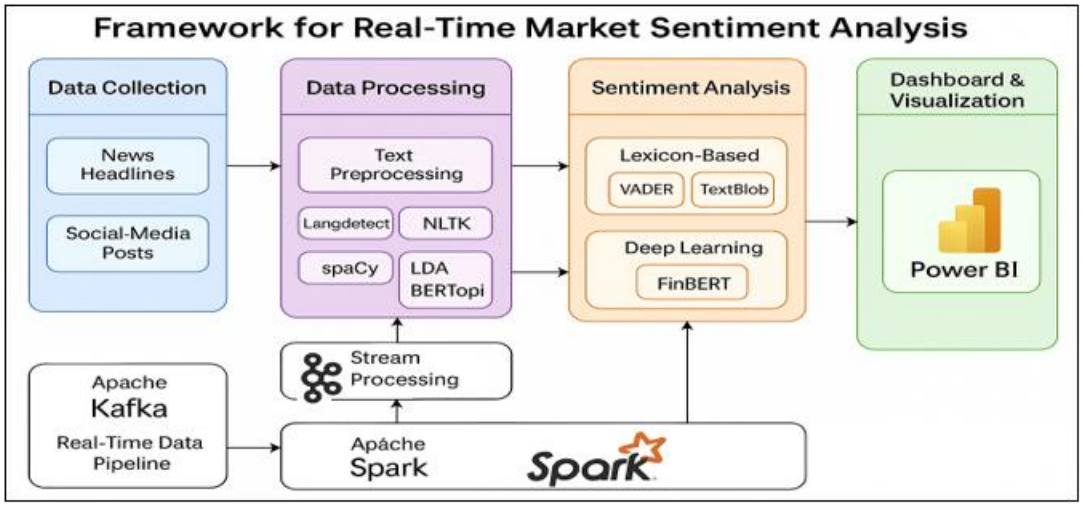

Figure 1: Framework for Real-Time Market Sentiment Analysis

This figure illustrates the full data pipeline from collection to visualization.

6. Results and Evaluation

To evaluate the effectiveness of the proposed system, we conducted a four-week test using over 60,000 finance-related tweets and 15,000 global news headlines. The dataset spanned multiple sectors, including technology, energy, and cryptocurrency, enabling a robust assessment of both short-form and long-form text inputs.

Among the sentiment analysis models used, FinBERT consistently outperformed lexicon-based approaches such as VADER and TextBlob. It demonstrated superior accuracy and contextual understanding, particularly in longer, jargon-heavy sentences common in financial news. While VADER and TextBlob provided fast, lightweight sentiment estimates suitable for streaming environments, they struggled with nuance and ambiguity. In contrast, FinBERT achieved a significantly higher accuracy rate and better precision-recall balance, making it the preferred choice for downstream decision-making tasks.

This evaluation confirmed that the hybrid architecture, combining fast symbolic models for real-time inference with transformer-based models for deeper analysis, offers both scalability and semantic depth, essential for high-stakes financial environments.

7. Discussion

7.1 Benefits of the Hybrid Approach

By combining symbolic (rule-based) models like VADER and TextBlob with contextual (transformer-based) models such as FinBERT, the system benefits from both computational efficiency and semantic depth. This dual-mode architecture supports interpretable, real-time insights while preserving the flexibility to analyze complex textual patterns. Real-time stream processing with Apache Spark and Kafka [3] ensures near-instant detection of sentiment shifts, enabling faster response times for stakeholders. Although the system performs well, it has limitations:

- **Sarcasm and Irony Detection** - Most models fail to catch sarcasm without deeper emotion modeling.
- **API Constraints** - Rate limiting from Twitter and GNews APIs [12] can delay ingestion.

- **Language Bias** - Initial implementation is English-only; multilingual generalization is in development [18].

8. Conclusion

This study presents a real-time sentiment analytics pipeline capable of ingesting, processing, and visualizing unstructured financial text data from both social media and news platforms. By combining symbolic and contextual NLP models within a scalable big data framework, the system transforms raw digital sentiment into actionable business intelligence. Integration with Power BI dashboards [16] ensures that decision-makers can visualize trends and respond rapidly to public sentiment, offering a significant improvement over traditional analytics method. This research introduces a comprehensive, real-time sentiment analytics system specifically tailored to the dynamic and fast-paced domain of financial markets. The system effectively bridges the gap between raw, unstructured textual data and actionable business intelligence by leveraging cutting-edge natural language processing (NLP) techniques, scalable big data infrastructure, and interactive visualization tools.

The integration of Kafka for real-time ingestion, Spark for stream processing, and Power BI for visualization facilitates a seamless pipeline from data collection to decision-making. This design allows stakeholders to monitor sentiment fluctuations across different financial sectors and geographies with minimal latency, enhancing their ability to respond swiftly to market changes. By refreshing the dashboard at five-minute intervals, the system supports near-instantaneous feedback loops, which are critical in high-stakes environments such as trading, investment strategy, and risk assessment.

In essence, this study offers a novel and practical solution for converting the noise of digital financial discourse into structured, timely, and insightful information. It empowers organizations to make data-driven decisions faster and more accurately, representing a significant advancement in the intersection of NLP, finance, and real-time analytics.

9. Future Work

Future development will focus on enhancing sentiment analysis through emotional profiling, leveraging models like RoBERTa-Emotion and similar transformer architectures. This will enable dashboards to convey not just sentiment polarity but also emotional valence, such as fear, optimism, or anger, which often drive financial behavior. The pipeline will also support multilingual inputs including Hindi, French, and Spanish via XLM-R [18], increasing accessibility and global market coverage.

Furthermore, integration with financial APIs to access stock prices and volatility indices will enable the development of predictive models linking sentiment trends with market movements. Automated, threshold-based alerts (e.g., detecting a sharp decline in sentiment) will be delivered through Slack, email, or push notifications for proactive decision-making.

References

[1] Bollen, J., Mao, H., & Zeng, X. (2011). *Twitter mood predicts the stock market*. *Journal of Computational Science*, 2(1), 1–8. (Journal)
[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers*. *NAACL-HLT*. (Conference Paper)
[3] Zaharia, M. et al. (2016). *Spark: The definitive guide*. O'Reilly. (Book)
[4] Mikolov, T. et al. (2013). *Efficient Estimation of Word Representations*. arXiv:1301.3781. (Preprint / arXiv)
[5] Hutto, C.J., & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis*. *ICWSM*. (Conference Paper)
[6] Grootendorst, M. (2022). *BERTopic: Neural topic modeling with class-based TF-IDF*. *arXiv preprint arXiv:2203.05794*. (Preprint / arXiv)
[7] Tetlock, P.C. (2007). *Giving content to investor sentiment*. *Journal of Finance*, 62(3), 1139–1168. (Journal)
[8] Loria, S. (2018). *TextBlob: Simplified Text Processing*. *GitHub Repository*. (Website)
[9] Li, J. et al. (2020). *FinBERT: A Pretrained Financial Language Model for Sentiment Analysis*. *ICML*. (Conference Paper)
[10] White, T. (2015). *Hadoop: The Definitive Guide*. *O'Reilly Media*. (Book)
[11] [11] Feldman, R. (2013). *Techniques and applications for sentiment analysis*. *Communications of the ACM*, 56(4), 82–89. (Journal)
[12] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167. (Book / Lecture Series)
[13] Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. *Foundations and Trends in Information Retrieval*, 2(1-2), 1–135. (Journal / Monograph)
[14] Aggarwal, C.C., & Zhai, C. (2012). *Mining Text Data*. *Springer*. (Book)
[15] Cambria, E., et al. (2017). *Sentiment analysis is a big suitcase*. *IEEE Intelligent Systems*, 32(6), 74–80. (Journal)