EEL 4798 - Massive Storage & Big Data

University of Central Florida

Project Proposal for

Real-Time Stock Market Analysis Using Big Data & Social Media Trends

By Colin Kirby

co201115@ucf.edu

February 4th, 2025

Abstract

The stock market generates vast amounts of structured financial data, while social media sentiment plays an increasing prevalent role in influencing trading decisions. However, traditional stock market analysis relies on batch processing, which can introduce delays and inefficiencies that limit decision-making. This research proposes a real-time stock market analysis pipeline that utilizes technologies like Apache Spark Streaming for fast data processing and Natural Language Processing (NLP) models to extract sentiment from platforms like Twitter and Reddit. By integrating various machine learning models, the system will predict stock movements using both financial and sentimental data, offering traders timely and data-driven insights. The system's performance will be evaluated based on latency, accuracy, and scalability, ensuring it can handle the demands of high-frequency trading. Future enhancements may include integrating news sources to further refine predictions and capture broader market trends.

Introduction

This project is an independent research study for EEL 4798 – Massive Storage and Big Data at the University of Central Florida. The assignment requires students to select a topic related to data-intensive computing and conduct a research-based study. As this is an individual project, all research, experimentation, and implementation will be conducted solely by myself. The selected topic focuses on real-time stock market analysis using big data and social media influence, exploring how large-scale financial data and unstructured sentiment from social media impact stock price movements. This study examines existing approaches, identifies their limitations, and proposes a real-time analysis framework designed to improve data processing efficiency, predictive accuracy, and scalability.

Context

The stock market is a complex, rapidly changing environment influenced by financial transactions, global economic events, and investor sentiment. Traditionally, stock analysis relied on historical price movements and technical indicators, but recent advancements show that social media sentiment also plays a significant role in market fluctuations. Platforms such as Twitter, Reddit, and StockTwits have become influential sources of stock-related discussions, offering real-time signals of market sentiment. However, extracting meaningful insights from these unstructured and often noisy data sources remains a challenge. To remain competitive, traders and institutions must integrate both financial and social media data into their decision-making process.

Problem Definition

Current stock market prediction models face major limitations in providing real-time insights due to challenges in latency, data quality, and scalability. The existing and commonly used batch processing methods introduce delays, making them unsuitable for high-frequency trading, where immediate insights are necessary for decision-making. Furthermore, social media sentiment analysis presents difficulties due to the presence of spam, fake news, and sarcasm, which make it challenging to accurately extract meaningful financial sentiment amongst the millions of posts. Furthermore, processing both structured financial data and unstructured social media data at scale requires significant computational power, which traditional models often struggle to handle efficiently. These challenges collectively lead to reduced prediction accuracy and increased market response times, highlighting the need for a real-time, scalable stock analysis solution that can seamlessly integrate financial data with sentiment-driven market trends.

Related Works

Stock market prediction has been a widely researched topic in the past, with more recent studies focusing on integrating big data, machine learning, and social media sentiment analysis to improve forecast accuracy. Previous research highlights the importance of real-time processing and scalable architectures as data of value will be active and able to handle expansions in dataset sizes. Challenges persist in extracting meaningful insights from this unstructured data and properly handling vast amounts of market info. This section will review some of the key studies that have informed our research, which will help us identify both effective methodologies and limitations in current approaches.

A study by Tsui (Stanford University) analyzed how aggregated StockTwits messages impact stock prices using machine learning models. This research found that bullish sentiment on social media often had a correlation with short-term price anomalies and that message volume spikes frequently signal abnormal price action. However, this study also highlights issues such as overfitting, frequent statistical inaccuracies, and selection bias when relying on social sentiment for stock predictions. This find supports our idea that while social media provides valuable insight, it should be used together with other financial indicators for more reliable predictions.

Another study, Social Media & Stock Market Prediction: A Big Data Approach, examined the role of Apache Spark MLlib in real-time stock price forecasting. The authors implemented liner regression, random forest, and generalized liner regression models, achieving an 80-98% accuracy range in predicting market trends. Additionally, Naïve Bayes and logistic regression models were tested for sentiment classifications, producing moderate success rates between 60-80% accuracy. The research has concluded that Spark MLlib significantly outperforms traditional batch processing models, reinforcing the need for real-time data streaming and machine learning integration in this realm of financial analysis. However, challenges in data volume management and feature selection were noted in the research, as determining meaningful stock indicators while filtering noise from social media remains an ongoing issue.

Further research by Awan et al. explored the impact of financial sentiment analysis from Twitter, Yahoo Finance, and Reddit on stock price movement. The study confirmed that the social media discussions influence market trends, yet prediction accuracy has remained inconsistent due to misinformation, sarcasm, and high data variability. The authors have found that sentiment polarity (bullish / bearish indicators) alone is insufficient, and then combining sentiment analysis with financial indicators leads to more reliable stock movement forecasting. The research also demonstrated that machine learning models such as Random Forest and LSTM outperformed traditional methods when trained on both numerical stock data and textural sentiment analysis. However, the paper identified significant challenges in scalability with these models, as real-time processing requires high performance computing resources to efficiently analyze multiple data streams simultaneously.

All these studies collectively highlight potential benefits and limitations of integrating big data analytics and social media sentiment analysis into stock market prediction. They demonstrate that while machine learning and sentiment classification models improve forecasting, noisy unstructured data introduces challenges that need to be addressed when approaching this topic. Our research will build upon this by proposing a real-time, scalable, stock market analysis framework that utilizes Apache Spark Streaming for continuous data ingestion, NLP models for improved sentiment extraction, and machine learning algorithms for predictive analytics. This approach seeks to overcome the scalability and accuracy limitations observed in previous research while ensuring low-latency financial data processing for high-frequency trading applications.

Proposed Solution

Now building on these findings from past research, this study proposes a real-time stock market framework that integrates big data processing, social media sentiment analysis, and machine learning-based predictive modeling. Unlike batch processing models, which introduce latency issues as found in our research, this approach utilizes Apache Spark Streaming to process the stock market data and social media sentiment in real time. And by incorporating Natural Language Processing (NLP) models, the system can more effectively filter out spam, fake news, and sarcasm, improving sentiment classification accuracy. Additionally, some machine learning modes such as Random Forest and LSTM will be used for analyzing the stock price trends by combining the financial indicators and sentiment data which will thus help improve predictive accuracy over standalone methods. This framework addresses the scalability challenges identified in previous studies by using a distributed computing design. While true distributed clusters typically run across multiple machines, this project will attempt to simulate a distributed environment utilizing Virtual Machines to emulate multi-node architectures. The approach enables scalable, and fault tolerant processing observed from past research while ensuring a low-latency financial processing application.

Experimental Methodology

To properly evaluate the effectiveness of our proposed solution's framework, this study will conduct a series of experiments focusing on latency, prediction accuracy, and scalability. These experiments will assess how well the system processes high-frequency stock data, extracts meaningful sentiment from social media, and generates accurate stock movements predictions compared to traditional batch-processing models.

The main data pipeline will integrate two primary data sources : financial market data from Yahoo Finance API and social sentiment data from platforms like Twitter or Reddit. Stock price trends will be processed in real time using Apache Spark, while NLP models will extract sentiment scores from social media posts to identify market sentiment trends. The machine learning models, including Random Forest, will be trained on the combined financial and sentiment data to predict stock price fluctuations.

To simulate our distributed computing environment, we will be using Apache Spark in Standalone Mode, while utilizing multiple CPU cores for proper parallel processing, as for this project multiple machines are not available, so everything will be run on a single machine. Additionally, the use of Virtual Machines may be used to test the system's ability to handle large-scale data loads under high-frequency trading conditions. Performance will be evaluated using key metrics like latency (our processing speed), prediction accuracy (correlation with actual market movements), and system scalability (ability to increase amount of data without a notable performance decrease).

By comparing these real-time predictions generated by our system with the actual market outcomes, the study will determine whether the proposed framework improves trading insights and decision-making speed over existing batch-processing methods. The findings from these experiments will help validate the potential of real-time big data processing and machine learning driven sentiment analysis in financial markets.

Expected Outcomes

Based on some of the findings from the previous works we researched, our stock market analysis framework is expected to improve upon prediction speed, overall accuracy, and general scalability combined to past and more traditional batch-processing methods. Prior research demonstrated that machine learning models using sentiment analysis improved stock movement predictions with accuracy rates ranging between 80 – 98% when using specific models like Random Forest or LSTMs. However, these studies did note challenges in handling more large-scale data efficiently and with that handling the noisy social media sentiment as well. By using a more "real-time" and refined sentiment approach with Apache Spark and NLP models, we are anticipating that we can reduce latency issues while still enhancing prediction accuracy.

Our approach to distributed computing will allow us to evaluate the system’s scalability under increasing data loads. Previous research detailed how Spark MLlib outperformed batch-processing models in handling financial data, and we are expecting similar gains in efficiency when applied to more high-frequency data. Ultimately, our system aims to improve upon existing research and better help traders gain access to more timely insights, and improved sentiment driven forecasting .

Conclusion & Future Work

This study aims to develop a real-time stock market analysis framework that improves accuracy, speed, and scalability by integrating financial and sentiment data. By addressing past limitations from previous works, this approach seeks to provide more reliable financial insights for use within the stock market. The experimental results will help determine the effectiveness of our approach to forecasting. Future works may be able to explore from the foundation of this work such as expanding data sources or refining predictive models to further enhance decision making in trading environments.

References

[1] D. Tsui, “Predicting Stock Price Movement Using Social Media Analysis,” *Stanford University*, 2016.

[2] M. J. Awan, M. S. M. Rahim, H. Nobanee, A. Munawar, A. Yasin, and A. M. Zain, “Social Media and Stock Market Prediction: A Big Data Approach,” *Computers, Materials & Continua*, vol. 67, no. 2, pp. 2570–2583, 2021.

[3] Huina Mao, Scott Counts, and Johan Bollen, “Quantifying the Effects of Online Bullishness on International Financial Markets,” *European Central Bank*, 2015.