PROJECT PROPOSAL

DATA 228 - BIG DATA TECHNOLOGIES

Stock Market Sentiment Analysis System

SAN JOSE STATE UNIVERSITY

GROUP - 10
AAYUSHI SHAH - 018180136
APURVA MUCHANDI - 018205785
DHRUVA JOSHI - 018191575
ROHIT DHOLE - 018195320

1. Abstract:

The real-time Stock Market Sentiment Analysis System is designed to provide a technological solution that helps retail investors make more informed decisions by analyzing the emotional tone and sentiment of financial news and social media content. The system operates in real-time, using *Kafka* to ingest live data feeds from multiple sources, including news articles, social media platforms, and financial reports. This data is then processed using *Spark Streaming*, which allows for the continuous and efficient analysis of large datasets.

The core of the system lies in its ability to analyze sentiment and perform entity recognition—identifying important market-related entities such as companies, stocks, and financial terms. *Sentiment analysis* algorithms classify news and social media posts as either positive, negative, or neutral, providing insights into how investors might be reacting to current events. The system then performs *correlation analysis* between the sentiment scores and stock market movements, allowing it to detect patterns and predict future market trends based on the prevailing sentiment.

The project uses three key datasets:

- **Financial News Dataset**: A collection of news articles related to stock markets, which is used to assess how market-moving news impacts sentiment.
- **Financial PhraseBanks**: A repository of financial phrases that can be used to better train sentiment analysis models on domain-specific language.
- **Stock Market Data**: Historical and real-time stock prices, which are used to correlate sentiment with actual market movements, allowing for predictions.

The outcome of this project is a robust and scalable system that can continuously ingest data, perform real-time sentiment analysis, and provide retail investors with insights into stock market movements based on the prevailing sentiment.

2. Motivation for the Project:

In financial markets, sentiment analysis has become a crucial tool for understanding market behavior and making predictions. Institutional investors typically have access to sophisticated tools and resources, such as high-frequency trading algorithms, proprietary sentiment analysis models, and large-scale financial data systems. These tools enable them to process vast amounts of data quickly and gain a competitive edge in predicting market movements. On the other hand, retail investors, who typically lack these resources, often struggle to interpret market data and news with the same level of sophistication.

This project aims to level the playing field by developing a real-time sentiment analysis system that can process financial news and social media content to provide insights into stock market movements. By incorporating machine learning models and natural language processing (NLP) techniques, the system empowers retail investors to gauge market sentiment, even without advanced tools. This is particularly important in a world where market sentiment can be influenced not only by traditional financial news but also by social media posts, blogs, and tweets.

Moreover, the project addresses several challenges faced by retail investors:

1. **Market Overload**: The volume of financial news and social media content makes it difficult for individuals to filter out meaningful insights from noise. By automating the sentiment analysis process, this system helps investors focus on actionable information.
2. **Reaction Time**: Traditional market research often lags behind real-time events. With real-time data ingestion and processing, the system ensures that investors can make timely decisions based on the latest market developments.
3. **Data Accessibility**: Retail investors often lack access to high-quality data and analytical tools. This system democratizes access to sentiment analysis, giving retail investors a more level playing field.
4. **Predictive Insights**: By correlating sentiment with stock price movements, the system not only tracks sentiment but also provides predictive insights that can guide investment decisions.

Ultimately, the motivation behind this project is to empower retail investors by offering a tool that combines big data technologies (Kafka, Spark Streaming) with advanced analytics (sentiment analysis and correlation with market movements) to produce actionable insights. This will help them understand the complex dynamics of the market and make more informed investment decisions, reducing the gap between retail and institutional investors.

3. Literature Survey

3.1 Sentiment Analysis in Financial Markets

Sentiment analysis has emerged as a critical tool in financial markets, where investor behavior is often driven by emotions and perceptions. Studies have shown that market sentiment, derived from news articles, social media, and financial reports, can significantly impact stock prices. For example:

- **Tetlock et al.** (2008) demonstrated that negative sentiment in news articles predicts downward pressure on stock prices, while positive sentiment correlates with price increases.
- **Bollen et al.** (2011) analyzed Twitter data and found that public mood, as reflected in tweets, could predict stock market movements with surprising accuracy.

However, most existing systems focus on either news or social media, leaving a gap in integrating multiple data sources for a more comprehensive view of market sentiment. Additionally, these systems are often proprietary and expensive, making them inaccessible to retail investors.

3.2 NLP for Financial Text

Financial text presents unique challenges for sentiment analysis due to its domain-specific language and context-dependent sentiment. For example:

- **Loughran and McDonald** (2011) developed a financial sentiment dictionary, showing that general-purpose sentiment analysis tools often fail to accurately capture sentiment in financial text.
- **Malo et al.** (2014) introduced the Financial PhraseBank, a dataset of annotated financial sentences, which has become a benchmark for training domain-specific sentiment analysis models.

These studies underscore the importance of fine-tuning NLP models for financial text, a key focus of this project.

3.3 Correlation Between Sentiment and Stock Prices

The relationship between sentiment and stock prices has been widely studied, but most research focuses on historical data rather than real-time analysis. For instance:

- **Sprenger et al.** (2014) found that Twitter sentiment could predict stock price movements, but their analysis was limited to historical data.
- **Li et al.** (2017) demonstrated the potential of real-time sentiment analysis for algorithmic trading, but their work was restricted to institutional investors.

Our project builds on these findings by developing a real-time system that integrates sentiment analysis with stock price correlation, making it accessible to retail investors.

4. Methodology

4.1 Data Collection and Preprocessing

Data Sources

- *Financial News Dataset:* A collection of 4+ million news articles from Kaggle, covering major US stocks from 2009-2020.
- *Financial PhraseBank:* A dataset of 5,000+ annotated financial sentences for training sentiment analysis models.
- *Stock Market Data:* Historical and real-time stock prices from Kaggle and Yahoo Finance.

Preprocessing

- *Text Cleaning:* Remove stopwords, punctuation, and special characters from news articles and social media posts.
- *Tokenization:* Split text into individual words or phrases for analysis.
- *Entity Recognition:* Use NLP techniques to identify companies, stocks, and financial terms mentioned in the text.

4.2 Sentiment Analysis Pipeline

Model Training

- *Fine-Tuning:* Will use the Financial PhraseBank to fine-tune a pre-trained NLP model (e.g., BERT) for financial sentiment analysis.
- *Sentiment Classification:* Classify each news article or social media post as positive, negative, or neutral.

Real-Time Processing

- *Kafka Integration:* Ingest real-time data from news APIs and social media platforms using Kafka.
- *Spark Streaming:* Process incoming data streams to generate sentiment scores in real-time.

4.3 Correlation Analysis

Time-Series Alignment

- Align sentiment scores with stock price data based on timestamps.
- Calculate lagged correlations to determine if sentiment predicts future price movements.

Anomaly Detection

- Identify unusual sentiment patterns (e.g., sudden spikes in negative sentiment) that may indicate market anomalies.
- Use statistical methods (e.g., Z-scores) to detect outliers.

4.4 Visualization and Insights

Dashboard Development

- Build a real-time dashboard using Tableau or Power BI to visualize sentiment trends and correlations.
- Include features like:
  - Real-time sentiment scores for selected stocks.
  - Historical sentiment trends.
  - Correlation heatmaps between sentiment and stock prices.

Backtesting

- Evaluate the system’s predictive power using historical data.
- Compare predicted price movements with actual market data to measure accuracy.

4.5 Evaluation Metrics

- *Accuracy:* Measure sentiment analysis accuracy against labeled data from the Financial PhraseBank.
- *Latency:* Evaluate real-time processing performance to ensure timely insights.
- *Correlation Strength:* Quantify the relationship between sentiment and stock prices using metrics like Pearson’s correlation coefficient.

5. Deliverables and Milestones

What Are We Delivering?

We’re building a real-time stock market sentiment analysis system that helps retail investors make smarter decisions by analyzing the mood of financial news and social media. Here’s what we’ll deliver:

1. A Working System:
    - ○ A tool that grabs live data from news and social media , processes it in real-time, and tells you whether the sentiment is positive, negative, or neutral. (Week 1-4)
    - ○ It also connects this sentiment to actual stock price movements, so you can see how people’s feelings about the market might affect stock prices. (Week 5-6)
2. Cool Visuals:
    - ○ We’ll create easy-to-understand charts and dashboards (using tools like Tableau or Python’s Matplotlib/Seaborn) that show how sentiment trends line up with stock prices. (Week 7-8)
3. A Detailed Report:
    - ○ A document that explains how we built the system, what we found, and how it can help retail investors. (Week 7-8)
4. Optional Research Paper:
    - ○ If time permits, we’ll write a shorter version of our findings and submit it to a journal or conference

The system will be developed over a full semester, with midterm deliverables including a functioning batch sentiment analysis pipeline and final deliverables including a complete real-time system with a user-friendly dashboard

6. Team Roles and Responsibilities

1. Data Collection and Ingestion- Dhruva

- ○ Set up Kafka to ingest real-time data from financial news and social media.
- ○ Integrate APIs like NewsAPI and Twitter API to feed data into the system.
- ○ Ensure data flows smoothly into the system without bottlenecks.

2. Data Processing and Sentiment Analysis- Aayushi

- ○ Use Spark Streaming to process the incoming data in real-time.
- ○ Develop sentiment analysis models using NLP libraries.
- ○ Perform entity recognition to identify key market-related terms

3. Correlation Analysis- Rohit

- ○ Analyze how sentiment trends correlate with stock price movements.
- ○ Use statistical methods to validate the strength of the correlations.

4. Visualization and Reporting- Apurva

- ○ Prepare Data for Visualization

- ○ Create interactive visualizations (e.g., dashboards, charts) using tools like Tableau or Python libraries.
- ○ Write the technical report summarizing the project's methodology, findings, and implications.

7. Relevance to the Course

How Does This Fit into the Course?
This project is perfectly aligned with what we're learning in class. Here's how:

1. Big Data Tools:
    - ○ We're using Kafka (for real-time data) and Spark Streaming (for processing), which are both covered in the course.
    - ○ This ties into CLO 3 (Using Tools Like Hadoop and Spark).
2. Real-Time Data:
    - ○ The project deals with live data from news and social media, which is all about handling volume and velocity—key topics in the course.
    - ○ This connects to CLO 8 (Transforming Data into Knowledge).
3. NLP and Sentiment Analysis:
    - ○ We're using NLP techniques (like sentiment analysis and entity recognition) to analyze financial text data. This ties into the course's focus on data analytics (Weeks 7-8).
    - ○ This aligns with CLO 4 (Designing and Implementing Solutions).
4. Predictive Insights:
    - ○ By connecting sentiment trends to stock prices, we're turning raw data into actionable insights. This is a big part of what the course is about.
    - ○ This fits with CLO 9 (Designing Analytical Solutions)

8. Technical Difficulty

- *Problems collecting data:* Real-time social media and financial data may be unstructured, noisy, or constrained by API restrictions.
- *Sentiment Analysis Challenges:* Misclassifications may result from unclear language, sarcasm, and financial jargon.
- *Complexity of Entity Recognition:* Because of name similarity, it is challenging to correctly identify businesses and financial phrases.
- *Limitations in Streaming:* When using Spark and Kafka for streaming to reduce latency, high-velocity data must be handled well.
- *Correlation and Prediction Problems:* While numerous external events impact markets, sentiment impacts on stock prices may be delayed.
- *Scalability Issues:* Real-time processing of massive datasets necessitates cost-effective cloud deployment and optimized infrastructure.
- *Dashboard Performance and Visualization:* Effective data synchronization is necessary for the real-time display of sentiment trends and correlations.
- *Overfitting Risks:* While models may function well on historical data, they may not be able to handle erratic market conditions.

9. Uniqueness

This project is unique in that it uses Kafka and Spark Streaming to provide real-time sentiment analysis, allowing for immediate insights from social media and financial news. In contrast to generic models, it uses Financial PhraseBank to refine BERT, guaranteeing precise interpretation of intricate financial terminology. Its capacity to find predictive patterns and associate sentiment trends with stock market movements is a crucial quality that aids investors in anticipating changes in the market. The solution bridges the gap between institutional traders and retail investors by democratizing access to real-time sentiment analytics and enabling data-driven decision-making.

10. Impact

1. An edge over competitors in market intelligence:

   By monitoring investor mood from social media and financial news, fintech companies can use this real-time sentiment research technology to improve market intelligence. In contrast to conventional approaches, it uses Kafka and Spark Streaming to deliver real-time insights, enabling businesses to keep up with swift changes in the market.

2. Improving Trading Strategies Through Algorithms

   FinTech companies can enhance their quantitative trading algorithms by combining sentiment-based signals with changes in stock prices. Identifying abnormalities in market sentiment allows for more intelligent trade execution, which lowers risk and boosts profitability.

3. Predicting volatility and managing risks

   Through the identification of abrupt changes in emotion that could affect stock prices, this approach helps in the early detection of market instability. FinTech businesses can use it to improve models for assessing portfolio risk and lessen losses brought on by sentiment-driven volatility.

To conclude, if successful, this project will level the playing field for retail investors, improve market transparency, and provide a scalable solution for sentiment analysis. It has the potential to transform how retail investors interact with financial markets, enabling them to make data-driven decisions with confidence.