Proposal on

Sentimetrics AI: Predicting Stock Prices through Sentiment Analysis

Submitted to

Department of Computer Science and Engineering

Nepal Engineering College

in Partial Fulfillment of the

Requirements for the Degree of B.E. in Computer

Submitted By

Sulav Timalsina (022-380)
Sushant Poudel (022-387)
Sushil Upadhyay (022-382)

Date: 03/12/2024

Abstract

The proposed project focuses on developing a comprehensive system for predicting stock prices by integrating sentiment analysis from social media and news articles with quantitative financial data. The system consists of four major modules: the Data Ingestion Layer, which collects data from various sources such as Twitter, Facebook, and financial APIs; the Data Processing & Analysis Layer, which conducts sentiment analysis and quantitative analysis using financial metrics and historical stock prices; the Prediction Engine, which merges insights from both analyses to forecast stock prices and includes a feedback loop for model improvement; and the User Interaction Layer, which provides insights and predictions through dashboards and APIs. The workflow involves gathering real-time data, preprocessing it, performing combined analyses, predicting stock prices, and continuously refining the model based on actual outcomes. Future enhancements will include adding more data sources and integrating macroeconomic factors to improve prediction accuracy and scalability.

Contents

Abbreviations

AI: Artificial Intelligence

NEPSE: Nepal Stock Exchange

BW: Behavior Weighted

EPS: Earnings Per Share

API: Application Programming Interface

BERT: Bidirectional Encoder Representations from Transformers

P/E: Price-To-Earnings

ARIMA: Autoregressive Integrated Moving Average

LSTM: Long Short-Term Memory

GraphQl:

List of Figures

Introduction

1.1. Overview of the system

The stock market is a complex ecosystem influenced by a multitude of factors, including economic indicators, company performance and increasingly public sentiment as expressed on social media. In this context, the project titled "Sentimetrics AI: Predicting Stock Prices through Sentiment Analysis" explores the dynamic relationship between social media sentiment and stock price movements of companies listed on the NEPSE.

With the rise of platforms like Twitter and Facebook, investors and stakeholders have started recognizing the power of public opinion in shaping market trends. However, the vast and unstructured nature of social media data makes it challenging to quantify its impact on stock prices effectively. Addressing this gap, the project introduces Sentimetrics AI, a cutting-edge artificial intelligence tool designed to analyze and predict the influence of social media sentiment on stock price fluctuations.

By leveraging advanced sentiment analysis techniques, Sentimetrics AI categorizes user comments views and news into positive, negative, or neutral sentiments. It then correlates these sentiments with stock price data, offering predictive insights that can assist investors in making informed decisions. This approach not only enhances the understanding of social sentiment and internet trends' role in market dynamics but also demonstrates the potential of integrating AI into financial analysis.

This project is significant in its ability to bridge the gap between social sentiment, internet trends and financial markets. It contributes to the emerging field of sentiment-based market analysis by providing a systematic framework to study the impact of online opinions on stock performance. In doing so, it offers practical applications for investors, analysts, and policymakers aiming to navigate the increasingly digital and sentiment-driven financial landscape.

Through this study, the project seeks to fill gaps in existing research by focusing on NEPSE, a relatively underexplored market, and provide actionable insights for both academics and practitioners in the field of finance and technology.

1.2. Problem Statement

Advanced sentiment analysis an extensive AI usage in prediction of NEPSE market is scarce. Sentimetrics AI aims to analyze public sentiments from social media posts and news as well as perform quantitate analysis of stock prices and tally their correlation with subsequent fluctuations in NEPSE prices, thereby enabling predictions of whether stock prices will rise or fall based on prevailing public sentiment.

1.3. Objectives

1.3.1. Main Objective

To analyze public sentiment on social media and news about NEPSE-listed companies as well as perform fundamental analysis and establish their correlation with stock price fluctuations for predictive insights.

1.3.2. Specific Objective

1. To analyze social media comments and views to classify them as positive, negative, or neutral sentiments related to NEPSE-listed companies.
2. To establish a correlation between the sentiment derived from social media and the stock price fluctuations of NEPSE companies.
3. To utilize Sentimetrics AI to predict stock price movements based on sentiment trends, providing actionable insights for investors.

Literature Review

This literature review examines existing research on sentiment analysis and its application in financial markets, particularly focusing on the correlation between social media sentiment and stock price movements.

NEPSE market sentiment analysis has emerged as a critical area of research, focusing on how investor sentiment influences stock prices, volatility, and overall market dynamics. The relationship between investor sentiment and NEPSE market behavior is complex and multifaceted, often characterized by both direct and indirect effects.

Research indicates that investor sentiment significantly impacts stock price volatility. For instance, [1] demonstrates that positive investor sentiment correlates with increased stock price volatility, while negative sentiment leads to reduced volatility. This finding aligns with the work of [5], who argue that divergence in sentiment among investors can lead to heightened price fluctuations, suggesting that sentiment-driven trading can influence market stability. Furthermore, the nonlinear dynamics of sentiment and stock returns have been explored by [2], who found that high sentiment tends to overvalue stocks, while low sentiment can lead to undervaluation, highlighting the asymmetric nature of sentiment effects on market performance.

The mechanisms through which sentiment affects the market are also noteworthy. [3] discuss the application of sentiment analysis techniques, including advanced models like transformers, to extract sentiment from large datasets. This sentiment extraction is crucial for forecasting stock market trends and making informed trading decisions. Additionally, the work of [6] emphasizes the importance of sentiment indicators, such as the BW composite sentiment index, which aggregates various market metrics to provide insights into investor behavior and its implications for stock performance.

In the context of technological advancements, the integration of sentiment analysis into financial decision-making has gained traction. The work of [4] introduces aspect-based

sentiment analysis, which allows for a more nuanced understanding of how specific sentiments related to different aspects of stocks can influence price movements. This methodological innovation enhances the ability to correlate sentiment with market dynamics effectively.

Many existing projects analyze the stock market using quantitative analysis rather than sentiment analysis itself [7]. A few focus on Sentiment Analysis. However, there is currently no similar study dedicated to the Nepalese stock market, known as NEPSE. We are referencing a similar project based on India [7].

In conclusion, the interplay between investor sentiment and stock market behavior is vital area of study that reveals how psychological factors can drive market volatility and influence investment strategies. The evidence suggests that sentiment not only affects immediate trading decisions but also has lasting implications for market stability and performance**.**

System Design

This section outlines the architecture and components of the proposed system for analyzing social media sentiment and predicting stock price movements.

3.1 System Architecture Overview

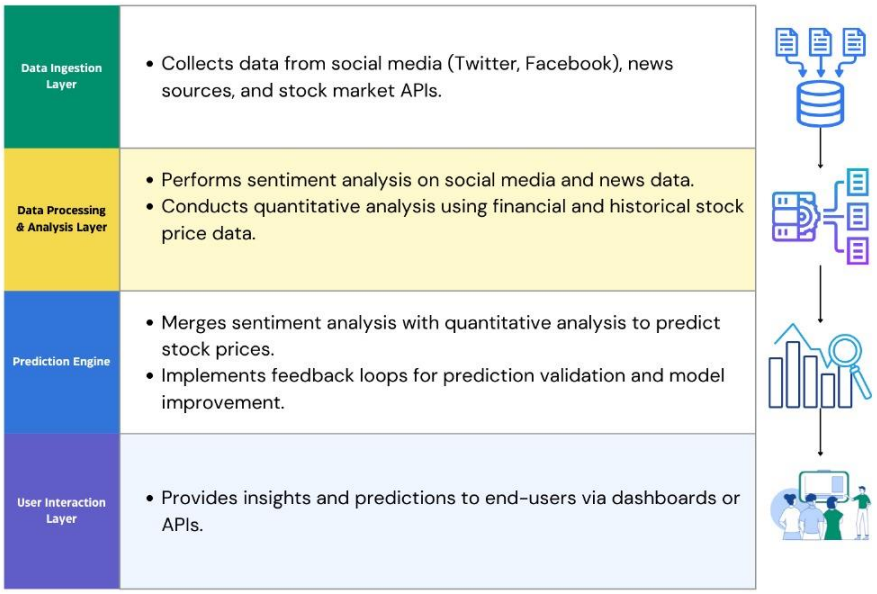

Figure 1- System Architecture Review

3.2 Detailed Design

3.2.1 Data Ingestion Layer

The Data Ingestion Layer is responsible for collecting data from various sources, including:

- Social Media: Utilize APIs such as the Twitter API and Facebook Graph API to gather real-time sentiment data.
- News Sources: Scrape relevant news websites or use APIs like NewsAPI to obtain articles and reports that may influence stock prices.
- Stock Market Data: Access financial APIs like Yahoo Finance or Alpha Vantage to retrieve historical stock prices and company financials.

Components:

- ETL Pipeline: Implement an Extract, Transform, Load (ETL) process to clean and normalize the incoming data.

- Message Queue: Use Apache Kafka or RabbitMQ to efficiently handle streaming data.

3.2.2 Data Processing & Analysis Layer

This layer performs two primary analyses: sentiment analysis of text data and quantitative analysis of financial metrics.

3.2.2.1 Sentiment Analysis:

- Preprocessing: Clean the text by removing noise (stopwords, links, emojis) and tokenize the text while handling language variations.
- Sentiment Model: Employ pretrained models like BERT or RoBERTa or train a custom model on labeled financial data.
- Output: Assign sentiment scores (positive, negative, neutral) to each piece of data.

3.2.2.2 Quantitative Analysis:

- Financial Metrics: Analyze key financial indicators such as revenue, earnings per share (EPS), opening and closing price, and price-to-earnings (P/E) ratios.
- Time-Series Analysis: Use models like ARIMA or LSTM to forecast trends based on historical stock prices.

3.2.3 Prediction Engine

The Prediction Engine integrates insights from both sentiment and quantitative analyses to make predictions about future stock prices.

- Model Integration: Combine the results of sentiment and quantitative analyses using ensemble models such as Random Forest or Gradient Boosting.
- Prediction Output: Generate predictions for stock prices over different time horizons (next day/week/month) along with confidence intervals.

- Feedback Loop: Continuously compare predicted prices with actual outcomes to refine the model using reinforcement learning techniques.

3.2.4 User Interaction Layer

This layer provides a user-friendly interface for stakeholders to access insights and predictions.

- Dashboard: Develop a visualization dashboard using frameworks like React or Angular to display sentiment trends, stock predictions, and model accuracy.
- API Development: Create RESTful or GraphQL APIs for integration with third-party applications and services.

3.2.5 Data Flow Diagram

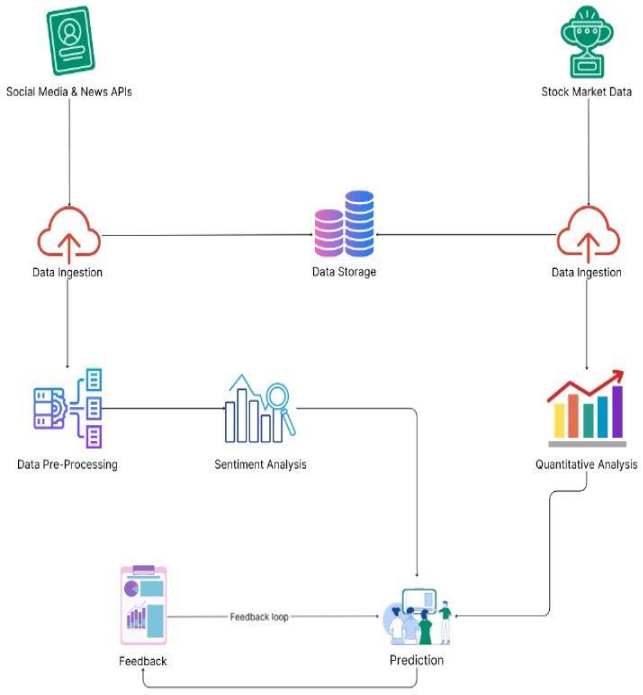

Figure 1:Showing Data Flow Diagram

3.3 Technology Stack

3.3.1 Data Ingestion:

- Python, Panda, NumPy, BeautifulSoup/Selenium, Facebook Graph API.

3.3.2 Data Processing & Sentiment Analysis:

- Natural Language Processing: NLTK, Hugging Face Transformers.
- ML Frameworks: TensorFlow, PyTorch.

3.3.3 Prediction Engine:

- Time Series Models: Statsmodels, Prophet.
- Ensemble Models: XGBoost, LightGBM.

3.3.4 Database:

- Relational: MySQL (for structured financial data).
- NoSQL: MongoDB (for unstructured social media/news data).

3.3.5 Deployment:

- Cloud Platforms: AWS, Azure, or GCP.
- Containerization: Docker, Kubernetes.

3.4 Workflow

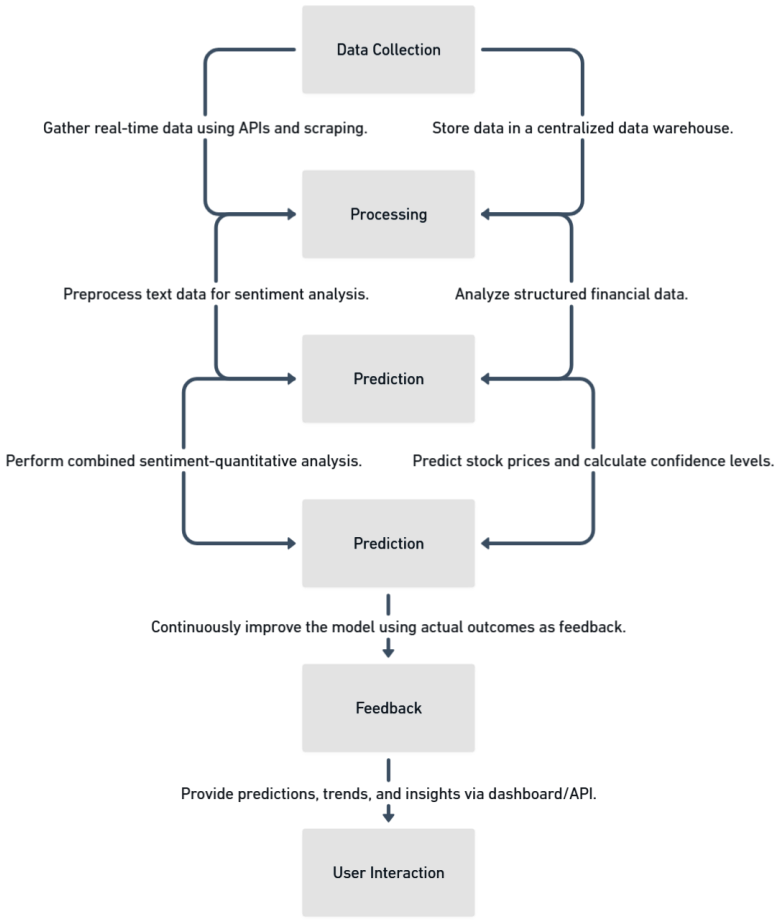

Figure 2: Showing Workflow Diagram

3.4.1 Data Collection:

- Gather real-time data using APIs and scraping.
- Store data in a centralized data warehouse.

3.4.2 Processing:

- Preprocess text data for sentiment analysis.
- Analyze structured financial data.

3.4.3 Prediction:

- Perform combined sentiment-quantitative analysis.
- Predict stock prices and calculate confidence levels.

3.4.4 Feedback:

- Continuously improve the model using actual outcomes as feedback.

3.4.5 User Interaction:

- Provide predictions, trends, and insights via dashboard/API.

3.5 Machine Learning Workflow

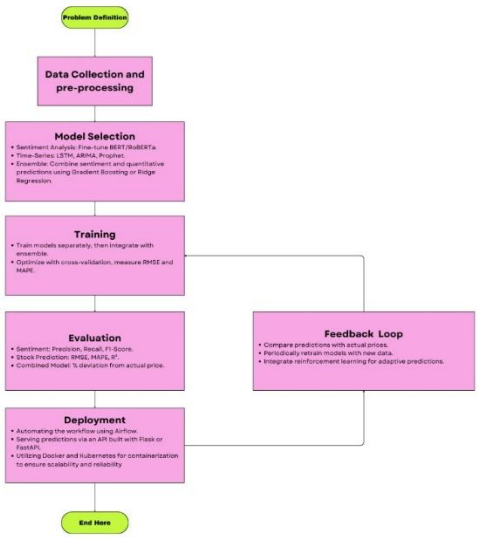

Figure 3: Showing Machine Learning Workflow diagram

3.5.1 Problem Definition

Predict NEPSE stock prices using sentiment analysis (social media, news) and quantitative analysis (financial data, historical prices).

3.5.2 Data Collection

- Social Media: Twitter API, Facebook API.
- News: NewsAPI, web scraping.
- Stock Data: NEPSE API, Yahoo Finance.

3.5.3 Data Preprocessing

- Text Data: Clean (remove noise), tokenize, lemmatize.
- Stock Data: Handle missing values, normalize, create rolling averages, volatility features.
- Aggregate: Merge sentiment and financial data with timestamps.

3.5.4 Feature Engineering

- Sentiment: Daily sentiment scores, trends.
- Financial: Historical prices, financial ratios, market indicators.
- Combined: Lagged sentiment and price features.

3.5.5 Model Selection

- Sentiment Analysis: Fine-tune BERT/RoBERTa.
- Time-Series: LSTM, ARIMA, Prophet.
- Ensemble: Combine sentiment and quantitative predictions using Gradient Boosting or Ridge Regression.

3.5.6 Training

- Train models separately, then integrate with ensemble.
- Optimize with cross-validation, measure RMSE and MAPE.

3.5.7 Evaluation

- Sentiment: Precision, Recall, F1-Score.
- Stock Prediction: RMSE, MAPE, R^2.
- Combined Model: % deviation from actual price.

3.5.8 Deployment

- Automating the workflow using Airflow.
- Serving predictions via an API built with Flask or FastAPI.
- Utilizing Docker and Kubernetes for containerization to ensure scalability and reliability

3.5.9 Feedback Loop

- Compare predictions with actual prices.
- Periodically retrain models with new data.
- Integrate reinforcement learning for adaptive predictions.

3.6 Scalability & Future Enhancements

- Add more data sources (Reddit, LinkedIn, etc.).
- Integrate macroeconomic factors like interest rates, inflation data.
- Improve real-time processing using advanced stream processing tools (e.g., Apache Flink).

Expected Output & Conclusion

The outcome of this project will be a comprehensive Sentimetrics AI Model, integrated with a fully functional software application featuring a user-friendly and visually appealing UI dashboard. The Sentimetrics AI Model will serve as the core analytical engine, capable of processing and classifying social media data into positive, negative, or neutral sentiments. It will include a predictive analytics module that correlates these sentiment trends with the stock price fluctuations of NEPSE-listed companies. Additionally, the model will leverage historical and real-time data to forecast potential market impacts, offering actionable insights to investors and stakeholders.

The software application will present this information through an intuitive dashboard designed for accessibility and ease of use. Key features of the dashboard will include visual representations of sentiment trends, stock price fluctuations, and the correlation between the two, allowing users to quickly grasp market dynamics. The dashboard will support customizable views, enabling users to filter data by company, time period, or sentiment type. Real-time updates will keep users informed of the latest market conditions, while predictive insights will provide forecasts of stock price movements based on the sentiment analysis.

Furthermore, the software will include advanced features such as sentiment heatmaps to visualize sentiment intensity across stocks or sectors, an alert system for significant sentiment changes or anomalies, and historical data analysis tools to explore long-term trends. Users will also have the option to export reports and visualizations in formats such as PDF or Excel for additional analysis.

This integrated solution will empower investors, market analysts, and researchers to understand the influence of social media sentiment on stock prices and make data-driven decisions. By combining cutting-edge AI technology with an intuitive interface, the project aims to bridge the gap between complex market analysis and practical application, fostering better-informed decision-making in the NEPSE stock market.

References

1. Guo, Q. (2023). The relationship between investor sentiment and stock market price. Frontiers in Business Economics and Management, 9(2), 124-129.

2. He, G., Zhu, S., & Gu, H. (2020). The nonlinear relationship between investor sentiment, stock return, and volatility. Discrete Dynamics in Nature and Society, 2020, 1-11.

3. Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. Ieee Access, 8,

4. Ong, K., Heever, W., Satapathy, R., Mengaldo, G., & Wang, Z. (2023).Finxabsa:explainable finance through aspect-based sentiment analysis..

5. Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2017). Divergence of sentiment and stock market trading. Journal of Banking & Finance, 78, 130-141.

6. Wang, X. (2022). The impact of investor sentiment on stock yield.

7. Gandalf1819, "Stock Market Sentiment Analysis," GitHub repository, 2024. [Online]. Available: https://github.com/gandalf1819/Stock-Market-Sentiment-Analysis. [Accessed: Dec. 5, 2024].