STOCK PRICE SIGNAL PREDICTION USING HISTORICAL DATA AND HYBRID INFORMATION

ARTIK GUPTA

Research Proposal

FEBRUARY 2025

ABSTRACT

Stock price prediction is always complex, considering the uncertainty of events that affect prices. However, the ability of AI to handle the complex task of detecting trends and patterns from historical data has helped ease the task of prediction to some extent. However, the latest algorithms and systems do not give high accuracy. Thus, there is a scope for innovation and research in this field. Leveraging the additional information such as public sentiments can further improve the accuracy.

This research paper delves into utilizing market sentiments with time series prediction algorithms. It delves into understanding various sentiment analysis techniques and feature creation. The paper will cover various aspects such as introduction, problem statement, research query, objectives, significance, scope, methodology, and necessary resources, along with final conclusion summary.

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
RNN	Recurrent Neural Network
SARIMA	Seasonal Autoregressive Integrated Moving Average
BERT	Bidirectional Encoder Representations from Transformers
AUC	Area Under Curve
API	Application Programming Interface

Table of Contents

1. Introduction

Stock price prediction is a complex problem for investors and traders. Correct prediction of stock price provides financial gain and is also a major driving factor of this research. Deciding the best price to buy or sell remains a challenging task for investors or traders because of price dependency on multiple factors including macro/micro economic factors and investor's sentiments. The market is influenced by macroeconomic policies, market sentiments and human psychology. Therefore, factors like social media and financial news can affect the stock price positively or vice-versa. Traditional stock price prediction methods are classified into two segments: technical and fundamental analysis. Technical analysis focussed on historical prices to predict future stock prices using statistical or graphical methods. On the other hand, the basic analysis relies on analysing financial information about a stock.

Stock price time series and other factors can be leveraged to tackle the problem of price prediction. This approach uses deep learning algorithms and sentiment variables extracted from financial news. Integration of sentiments or market factor variables will be beneficial in improving prediction accuracy. Machine learning algorithms can process large amounts of financial information which can be gathered using automated ways such as web scraping. Incorporation of financial information into technical analysis fills the prediction gap which remains by using only technical or fundamental analysis. Thus, this research aims to experiment with sentiment extraction from financial text information and will further use these sentiment scores as a feature in price prediction model.

2. Problem Statement

Stock price prediction is a relevant domain owing to its potential to generate financial value. Many data mining algorithms have been proposed to predict stock. However, historical information cannot predict the price the next day due to uncertain events that affect stock prices.

Traditional prediction methods are reliable for capturing regular, seasonal, and structured data. Researchers have leveraged various traditional models such as SVM, and ANN. Both of them are significantly effective for predicting stock price movement up to an extent. Furthermore, CNN and LSTM were widely used in prediction. (Jing et al., 2021)

According to the Efficient Market Hypothesis stock market prices are affected by financial information, i.e. news, and social media trends, rather than only present and past prices. As events affecting stock prices are unprecedented, stock market prices follow a random trend which cannot be determined with more than 50% accuracy. Thus, when an uncertain event occurs, the prediction results are unsatisfactory (Li et al., n.d., 2017). On the internet, investors express their views or sentiments through social platforms, financial websites and news articles, which impact the rationality of other investors and drive other investors' decision-making (Wu et al., 2022)

People express their views and comment on social media and news articles that others can see. Thus, generating a common sentiment among many investors. Therefore, analysing the financial text information related to a stock may help in classifying a sentiment about that stock. (Bharathi and Geetha, 2017). It has been observed that there is a strong correlation between news articles related to stocks and stock prices (Mohan et al., 2019)

Therefore, accuracy of the prediction model depends on the accuracy of sentiment analysis. Various research has been done to classify a sentiment form a text. For example, (Nemes and Kiss, 2021) used RNN and NLTK Lexicon architecture to predict stock prices with market sentiment information. (Joshi et al., n.d.) used a dictionary-based approach for tagging finance-specific sentiment words. They have used Random Forest and SVM to classify information. (Nallapati et al., 2016) proposed an interpretable neural model for document summarization that performs better than other deep learning models for summarizing documents. (Maqbool et al., 2022) used VADER from NLTK, TextBlob from NLP, and Flair from NLP to classify each news headline sentiment.

Nevertheless, current methods have limitations in terms of textual information classification and its utilisation in prediction. Market fluctuations due to extreme events (such as disasters, pandemics, wars, and droughts) and economic movements impact the stock price directly or indirectly. Furthermore, news data extracted from websites at times, led to news items which were not as relevant to the company in question. Moreover, strong positive or negative sentiments based on single news may give incorrect signal predictions. (Darapaneni et al., 2022)

Therefore, limitations include textual information classification and incorporating this information with price prediction deep learning or machine learning models. Addressing these problems is fundamental for improving the accuracy of the stock price prediction models.

3. Research Queries

The following research questions are suggested which are given below.

- How can financial text improve stock price prediction?
- What challenges exist in predicting stock prices during uncertain events?
- How can stock price prediction in uncertainty be further improved?

4. Aim & Objectives

The study will start with understanding the background of feature creation from textual information. It then delves into the existing methodologies and tools to calculate the significance of textual data and identifies their limitations. The research questions guiding this research aim to improve the accuracy of stock price prediction using filtered textual information from the internet.

The objectives are as follows:

- To explore existing techniques and methodologies for financial text classification.
- To understand the techniques used for forecasting stock price including Random Forest, ANN, and LSTM- RNN algorithms among others.
- To find the best possible way of incorporating financial information with prediction algorithms.
- To identify limitations and challenges associated with stock price prediction using historical time series data.

This research aspires to assess the use of textual financial information in stock price prediction. Specific objectives include measuring the accuracy of prediction and measuring the impact of external data on price prediction.

5. Significance

This study is significant in terms of financial value-generation capability. Achieving a high accuracy in price prediction can be beneficial in investor markets. Additionally, the outcomes of this research can help understand the limitations, challenges and possible solutions for price prediction.

6. Scope

This research will focus on a stock price prediction system based on sentiment analysis from financial information which includes news, social media reactions, and company fundamental information. The scope includes exploring various techniques for integrating textual information in deep learning-based prediction systems, the accuracy of the system in the prediction of price, the assessment of limitations, and the scope of improvement.

7. Methodology

The research methodology involves historical and textual data collection, data cleaning and validations, feature creation, price prediction model building, financial information integration, and assessment of the model’s accuracy. Various AI algorithms and datasets will be used to create models and assess accuracy. A comparative assessment will be done to assess the impact of sentiments on stock price. Figure 7.1 shows a basic architecture or roadmap for a stock price prediction model using sentiment analysis.

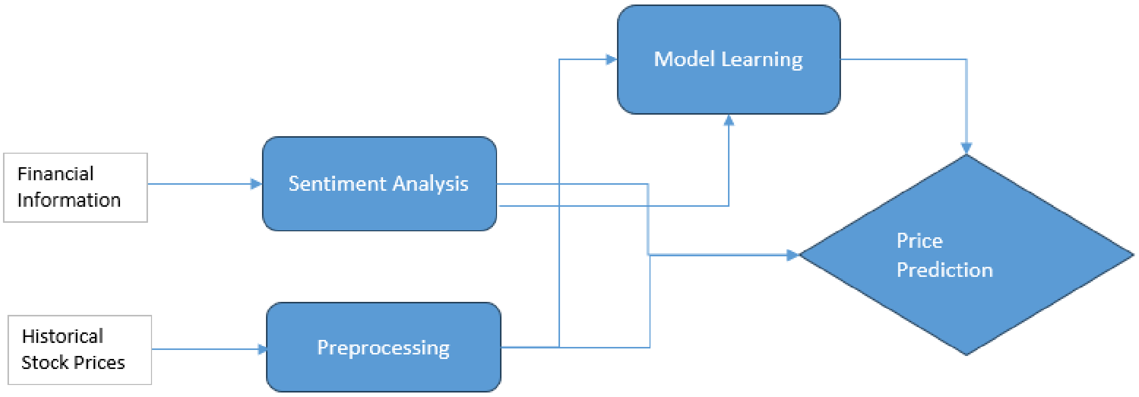


Figure 7.1: *Basic Architecture of Stock Price Prediction System*

The research methodology will involve the following steps:

- **Data Collection:** Collect historical stock price data, and scrape financial information including financial news, and social media trends related to stock for experimentation.
    1. **Data set:** The dataset to be used is: Adani Green Energy Limited (ADANIGREEN.NS) Stock Historical Prices & Data - Yahoo Finance, Scraped financial information from the web, supporting code snippets and libraries info.
    2. **Data pre-processing:** Data pre-processing includes the following steps:
    3. **Text Scraping**: Create codes for searching, filtering and scraping financial information related to stock from the web. It involves financial news extraction, and social media data scraping using their respective API.
    4. **Text cleaning and filtering**: Clean scraped text from the web by removing stop words or special characters that may not be recognized by language models. For example: special characters, string literals, and underscores need to be handled before using the text for prediction. Furthermore, the cleaned text needs to be filtered considering its importance in stock price prediction. Text encoders are to be leveraged in this process and text is to be filtered after getting high similarity scores with our desired keywords or sentences.
    5. **Historical Stock Price Data Cleaning**: Historical data has missing values due to holidays, and uncertain events. Thus, missing values need to be filled or tagged. Apart from that, outliers due to extreme events have to be handled effectively.

- **Exploratory data analysis**: EDA has to be done to analyse the training data. Univariate, Bivariate and multivariate analyses can be done to develop a deep understanding of the dataset.
- **Feature Creation:** Vectorization methods to be applied to extract meaningful information from financial information. A similarity score is to be calculated from the embedded data to create a feature representing public sentiment. Furthermore, other features including stock fundamentals, and historical prices, such as moving averages, rolling averages and lagged series among others will be created using feature engineering methods for financial purposes.
- **Sentiment Analysis**: Sentiment analysis has to be done to incorporate market sentiments with price prediction. Thus, each piece of information needs to be converted to numerical vectors using popular encoders. For example, BERT is a bidirectional model which captures deeper meaning by reading in both directions, not just

sequentially forward. This can be used to create sentence embeddings resulting in numerical vectors. As shown in Fig 7.2, a sentiment score or similarity score and a categorical variable are to be created to feed in the prediction model as a feature.

1. **Sentiment Analysis accuracy check**: Categorically classified financial information will be evaluated using various metrics including AUC, precision, recall, and F1 Score.

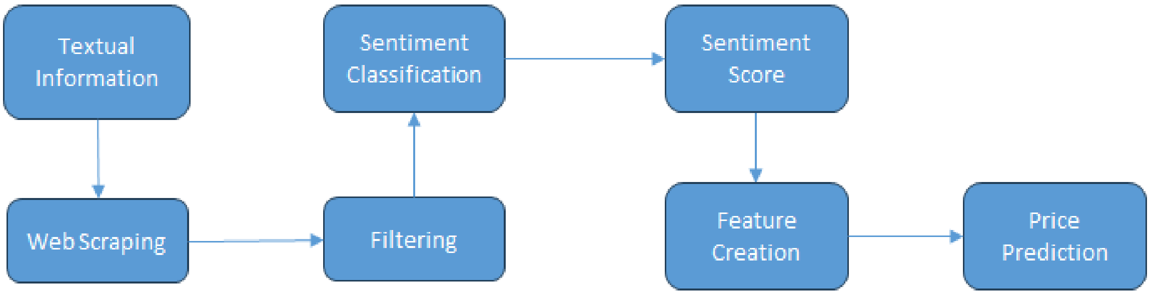


Figure 7.2: Sentiment Analysis Road Map

- **Prediction model**: There are various traditional prediction models such as Regression, SARIMA, ANN, and RNN (LSTM) among others. However, ANN and other models are unable to predict the uncertain effects in the stock market due to the absence of memory elements (Moghar and Hamiche, 2020). We will use the Long Short-Term Memory (LSTM) model based on the Recurrent Neural Network (RNN) which has memory cells that maintain information over long periods.
- **Code Generation**: Web scraping, stock data importing and cleaning, text data preprocessing, model creation and evaluation have to be done using python3 scripting language. Code creation steps are further elaborated below:
  1. **Syntax Generation**: Required logic and steps to be created at first as per requirement using various Python libraries including Pandas, NumPy, NLTK, Regex, BeautiSoap, HTML, sci-kit learn, SciPy, maths, and TensorFlow among others.
  2. **Model Training**: The model to be trained on train data which would be data from the past three or four years. Train data includes time series and scraped data from the internet. Test data would be the latest six months' time series data and financial data. There will be data for cross-validation to improve the model's accuracy and prevent overfitting.
  3. **Error and Exception Handling**: All if-else cases and handling of each error possibility are required.

4. **Visualization**: Time series data to be plotted using libraries such as matplotlib or Plotly to discover trends and patterns in stock price movement over time. The final model prediction will also be plotted against actual test data to visualise an error estimate. Other charts can also be plotted such as scatter plots to show stock price and sentiment analysis scores.

- **Evaluation:** Model evaluation will be done on two processes in the system.
  1. First, there is sentiment analysis. The score obtained from sentiment analysis has to be validated. Various methods, including K-Fold cross-validation, are used on a pre-classified textual dataset to perform validation.
  2. Second, is evaluating the final prediction model performance. This can be done using various metrics including:

     MAE: Mean Absolute error quantifies the average of absolute errors between predicted and actual values. It is used to check the accuracy of prediction models. A lower value indicates better accuracy and vice-versa.

     MSE: Mean square error is calculated by taking an average of squared differences between predicted and actual values. However, because of squaring, larger errors become more significant compared to lower errors.

     Root Mean Square Error: It handles MSE issues by converting numbers to their original units.

     MAPE: Mean Absolute Percentage Error expresses error as a percentage. A lower mape value indicates higher accuracy.

8. Required Resources

The research requires access to jupyter notebooks, python terminal, and ide for experimentation, and computing resources for implementation.

8.1 Requirements (Hardware)

- **Processor (CPU):** A multi-core processor (e.g., Intel i7 or i9, AMD Ryzen 7 or 9 will be required for efficient data processing and model training.
- **Memory (RAM):** At least 16GB of RAM, or more is preferable for handling large datasets and smooth performance.
- **Storage:** SSDs (Solid State Drives) will be needed for faster data access and processing.
- **Internet Connection:** A fast internet connection is required for real-time data retrieval from sources like Twitter, or financial news websites.

8.2 Requirements (Software)

Software and libraries include:

- **Deep learning frameworks:** TensorFlow, PyTorch, or similar frameworks for implementing and training stock prediction models.
- Language processing framework such as hugging face.
- IDE such as Visual Studio and Spyder to create prediction systems.

8.3 Dataset Specifications

There will be two datasets which include:

- stock's historical time series data about opening, closing prices and average prices rolled up on day, week, month or year level. Different stocks can be selected for to experiment.
- Financial sentiment information: This can be news or social media trends or other articles taken from the web and available in the public domain.

8.4 Model Training Details

Training a stock price prediction model will involve these steps:

- Data Scraping: Text data will be scraped and cleaned from all extra noise or stop words and characters.
- Pre-processing: Text data must be filtered using similarity scores before training.
- Model architecture: Choose an appropriate architecture for sentiment analysis and price prediction.
- Hyperparameters: Tuning hyperparameters such as learning rate, batch size, and dropout to improve the model's performance.
- Training procedure: Training includes improving accuracy by feature engineering, cross-validation, and parameter tuning.
- Evaluation: The model will be evaluated on a test data set from the latest period data.

9. Research Roadmap

A research plan spanning 23 weeks from January 14th, 2025, when research topics were approved:

Weeks 1-4: Research Proposal Development

- Weeks 1-2: define aim, scope and methodology.
- Weeks 3-4: Draft the research proposal, including the literature review and theoretical framework.

Weeks 5-6: Proposal Review and Revision

- Week 5: Submit the proposal for review.
- Week 6: Update the proposal based on feedback and submit it after finalization.

Weeks 7-10: Data Collection and Analysis

- Weeks 7-9: Collect data through web scraping and stock data apps such as Yahoo Finance. Clean and filter data to perform analysis.
- Week 10: Perform EDA and create proper visual and descriptive reports using jupyter notebooks and other tools.

Weeks 11-14: Prediction system creation and evaluation

- Weeks 11-12: Create the model's architecture, train-test model and evaluate.
- Weeks 13-14: Draft the results section of the report.

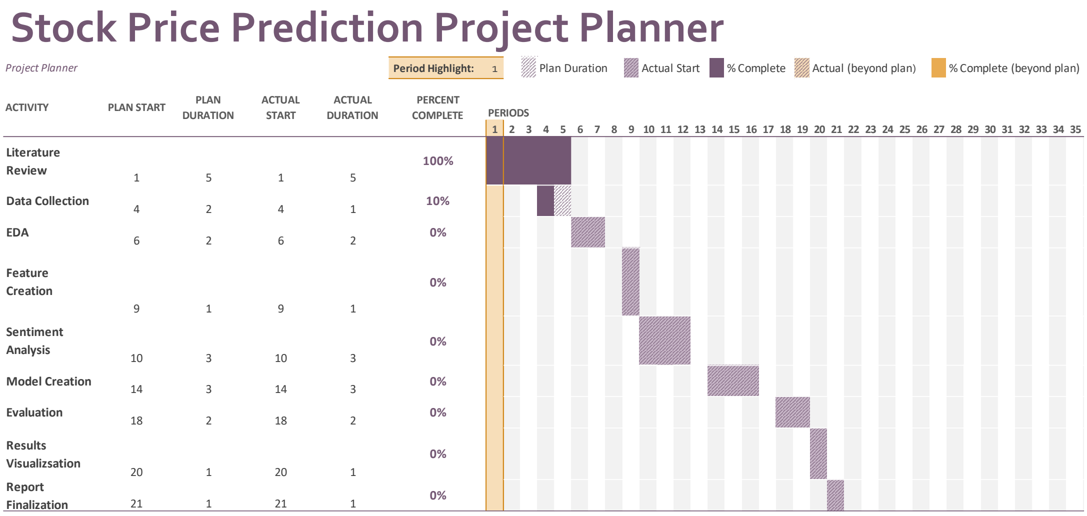
Weeks 15-18: Writing and Revision

- Weeks 15-16: Write conclusion sections having findings and a final summary.
- Weeks 17-18: Update the entire research report based on feedback.

Weeks 19-22: Presentation Preparation and Finalization

- Weeks 19-20: Prepare visual aids and presentation materials. Practice the presentation for conferences or seminars.
- Weeks 21-23: Finalize the research report, create a presentation and submit for assessment.

Here is the Gantt chart for the project timeline:



Stock Price Prediction Project Planner

Project Planner

Period Highlight: 1 ▨ Plan Duration ▨ Actual Start ■ % Complete ▨ Actual (beyond plan) ■ % Complete (beyond plan)

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Literature Review	1	5	1	5	100%
Data Collection	4	2	4	1	10%
EDA	6	2	6	2	0%
Feature Creation	9	1	9	1	0%
Sentiment Analysis	10	3	10	3	0%
Model Creation	14	3	14	3	0%
Evaluation	18	2	18	2	0%
Results Visualizsation	20	1	20	1	0%
Report Finalization	21	1	21	1	0%

Figure 9.1. Project Plan

This timeline offers a structured approach for completing the research process over 22 weeks. This will ensure sufficient time for proposal development, data collection, analysis, writing, presentation, and finalization.

References

Bharathi, S. and Geetha, A., (2017) Sentiment analysis for effective stock market prediction. *International Journal of Intelligent Engineering and Systems*, 103, pp.146–154.

Darapaneni, N., Reddy Paduri, A., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., Aiyer, U. and Agarwal, Y., (2022) *Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets*.

Jing, N., Wu, Z. and Wang, H., (2021) A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, 178.

Joshi, K., Bharathi, H.N. and Rao, J., (n.d.) *STOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS*.

Li, J., Bu, H. and Wu, J., (2017) *Sentiment-Aware Stock Market Prediction: A Deep Learning Method*.

Maqbool, J., Aggarwal, P., Kaur, R., Mittal, A. and Ganaie, I.A., (2022) Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. In: *Procedia Computer Science*. Elsevier B.V., pp.1067–1078.

Moghar, A. and Hamiche, M., (2020) Stock Market Prediction Using LSTM Recurrent Neural Network. In: *Procedia Computer Science*. Elsevier B.V., pp.1168–1173.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. and Anastasiu, D.C., (2019) Stock price prediction using news sentiment analysis. In: *Proceedings - 5th IEEE International Conference on Big Data Service and Applications, BigDataService 2019, Workshop on Big Data in Water Resources, Environment, and Hydraulic Engineering and Workshop on Medical, Healthcare, Using Big Data Technologies*. Institute of Electrical and Electronics Engineers Inc., pp.205–208.

Nallapati, R., Zhai, F. and Zhou, B., (2016) SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. [online] Available at: http://arxiv.org/abs/1611.04230.

Nemes, L. and Kiss, A., (2021) Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, 53, pp.375–394.

Wu, S., Liu, Y., Zou, Z. and Weng, T.H., (2022) S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. *Connection Science*, 341, pp.44–62.