FinClub Open Project 2024

Stock Sentiment Analysis Using Machine Learning

Bhawana Yadav | 21112036
IIT Roorkee
b_yadav@ch.iitr.ac.in

Problem Statement

The project aims to develop a sentiment analysis model to predict the movement of stock prices based on textual data from news articles, social media posts, and other sources of financial news and opinions. By analyzing the sentiment expressed in these texts, the model will seek to uncover insights into investor sentiment and market sentiment, which can be valuable indicators for making informed trading decisions.

Objectives: Collect a large dataset of textual data related to stocks, including news articles, social media posts, earnings reports, and analyst reports, from various sources.

Preprocess the textual data by removing noise, tokenizing text into words or phrases, and applying techniques such as stemming and lemmatization to standardize text representations.

Label the textual data with corresponding stock price movements (e.g., increase, decrease, or no change) over a specified time horizon to create a labeled dataset for supervised learning.

Extract features from the textual data, such as word frequencies, sentiment scores, and topic modeling representations, to represent the text in a format suitable for machine learning algorithms.

Train and evaluate machine learning models, including classification algorithms such as logistic regression, support vector machines (SVM), random forests, and neural networks, to predict stock price movements based on textual sentiment.

Perform model evaluation using appropriate metrics, such as accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curve analysis, to assess the performance of the sentiment analysis model.

INTRODUCTION

In the dynamic world of financial markets, making informed and timely investment decisions is crucial. Traditional financial analysis methods typically rely on quantitative data, such as historical price movements and financial ratios. However, the advent of digital media has brought forth a significant amount of qualitative information that provides valuable insights potentially influencing market behavior.

This project, titled "Stock Sentiment Analysis Using Machine Learning Techniques" (Project Code: #FC24OPS3), aims to leverage natural language processing (NLP) and machine learning to analyze textual data and predict stock price movements.

The fundamental idea of sentiment analysis in financial markets is to assess the mood or sentiment of investors and market participants as reflected in various textual sources. These sources include news articles, social media posts, earnings reports, and analyst opinions. The hypothesis is that the collective sentiment expressed in these texts can indicate future stock price movements. Positive sentiment may imply bullish behavior, whereas negative sentiment might suggest bearish trends.

By harnessing the power of NLP and machine learning, this project seeks to transform qualitative textual data into actionable insights for investors.

Implementation Strategy

The project follows a structured implementation strategy, divided into several phases, each focusing on a critical aspect of the sentiment analysis pipeline:

1. Data Collection and Preprocessing:

- Text Extraction: Utilize regular expressions to extract user mentions and hashtags from tweets.

- Contraction Expansion: Standardize text by expanding contractions to their full forms.

- Text Cleaning: Remove punctuation, special characters, URLs, and words with digits to ensure clean and consistent text data.

- Tokenization and Normalization: Break down text into individual tokens, remove stopwords, and apply lemmatization to reduce words to their base forms.

- POS Tagging and Chunking: Use Part-of-Speech (POS) tagging to identify grammatical parts of speech and custom chunking techniques to extract noun phrases.

- Sentiment Analysis: Compute sentiment scores using the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool.

2. Feature Extraction and Labeling:

- Feature Extraction: Derive features such as word frequencies, sentiment scores, and topic modeling representations from the text.

- Labeling: Annotate the textual data with corresponding stock price movements (increase, decrease, or no change) over a specified time frame.

3. Model Training and Evaluation:

- Classification Models: Employ several machine learning models to classify the data.

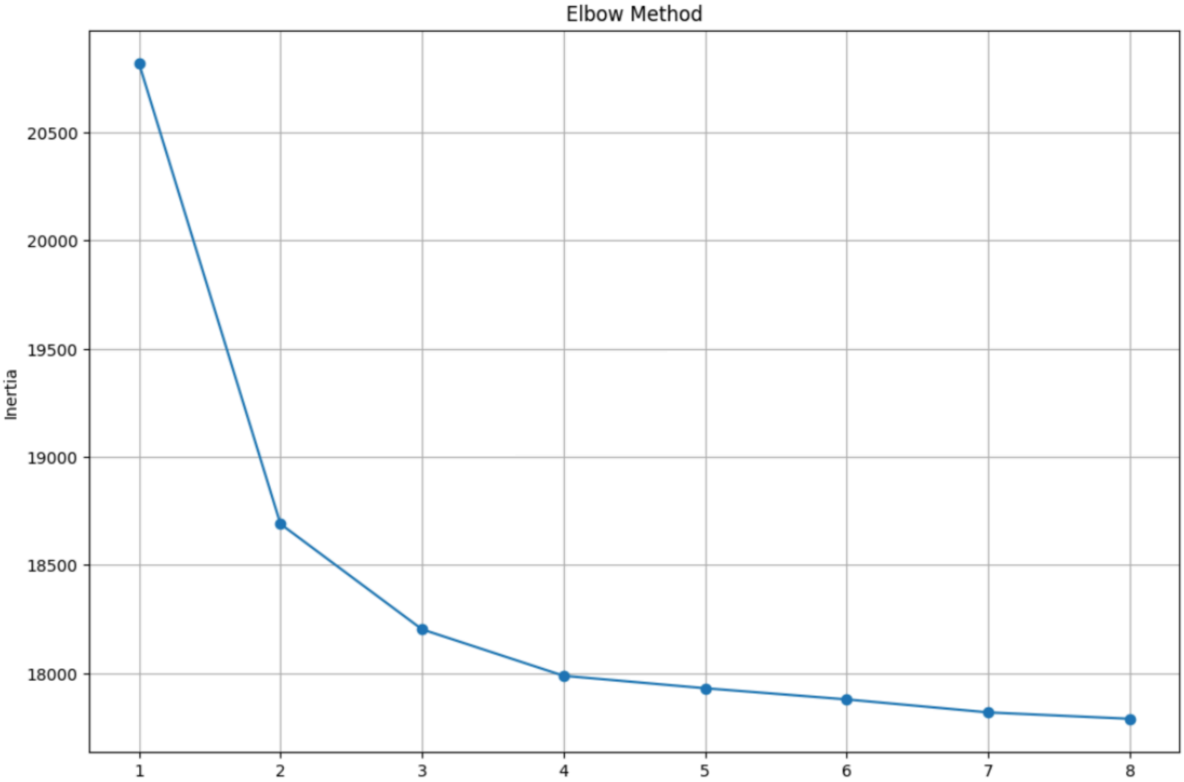
- Logistic Regression: A linear model ideal for binary classification tasks.

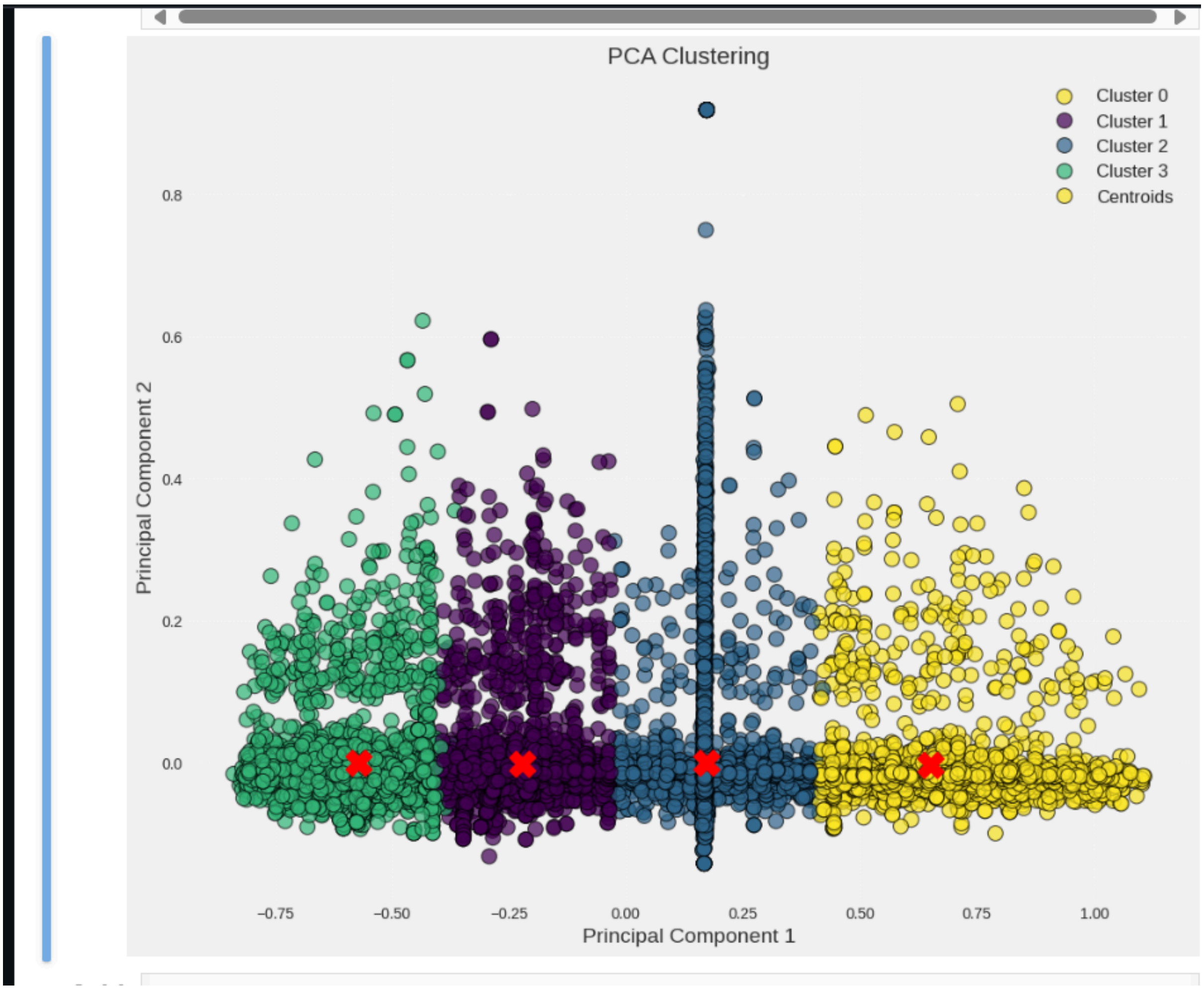
- Support Vector Machine (SVM): A model designed to identify the optimal hyperplane that best separates the different classes.

4 ROC AUC Score: The area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes

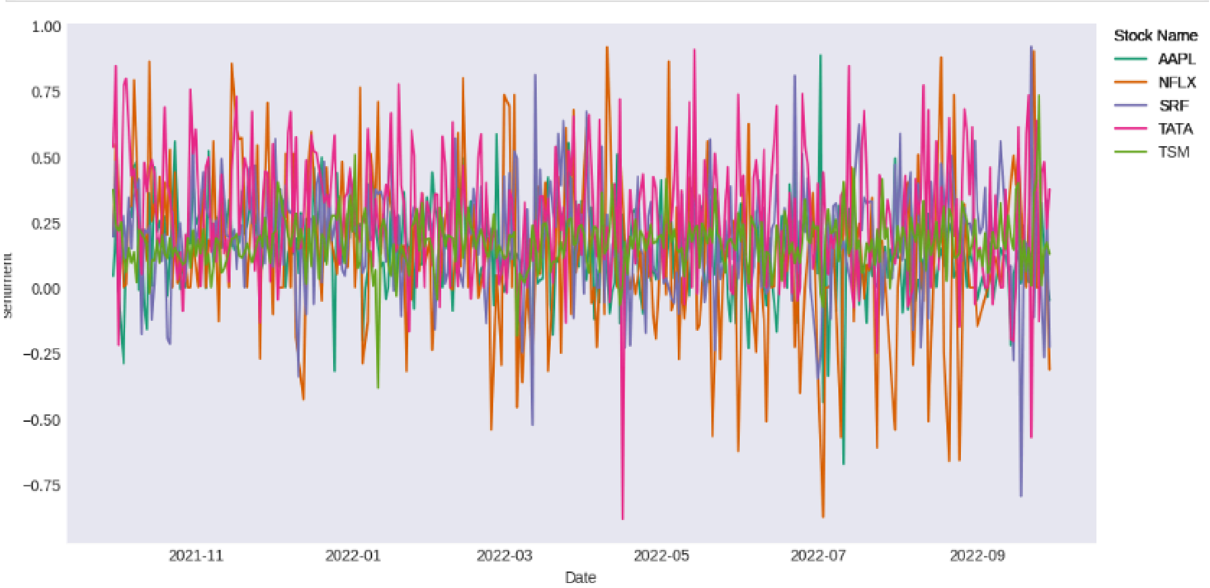
This structured approach ensures each phase is meticulously executed to leverage natural language processing and machine learning techniques for predicting stock price movements based on textual sentiment analysis.

Cluster -

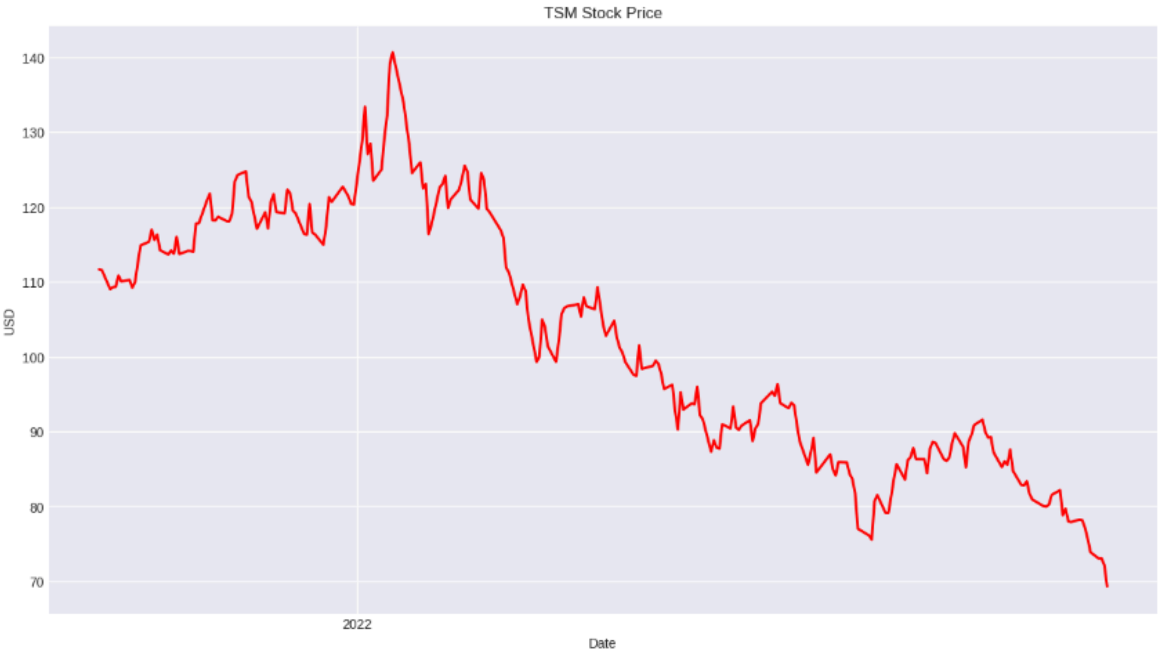

Elbow Method
Inertia
20500
20000
19500
19000
18500
18000
1
2
3
4
5
6
7
8



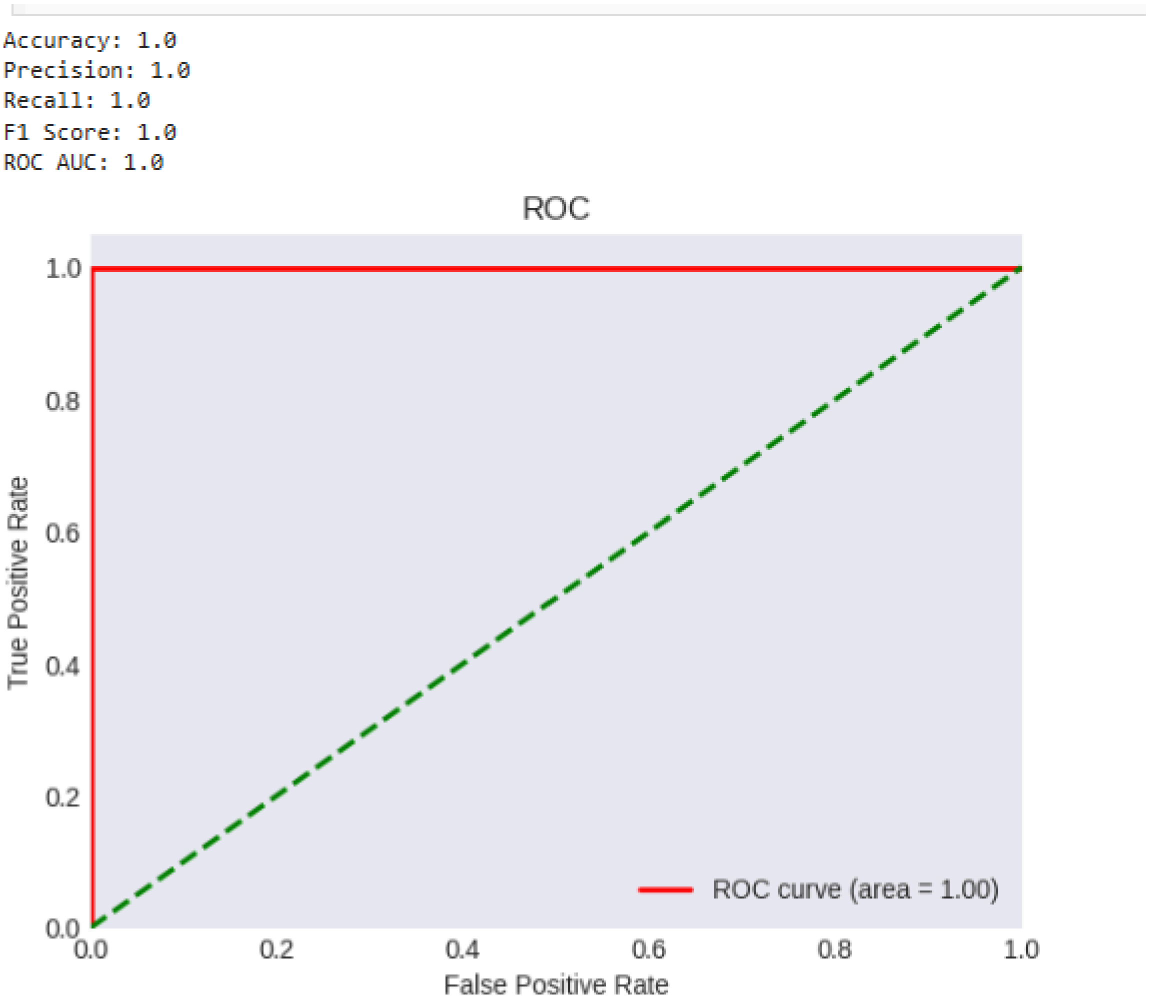
Sentiment Score-



TSM-Stock Price

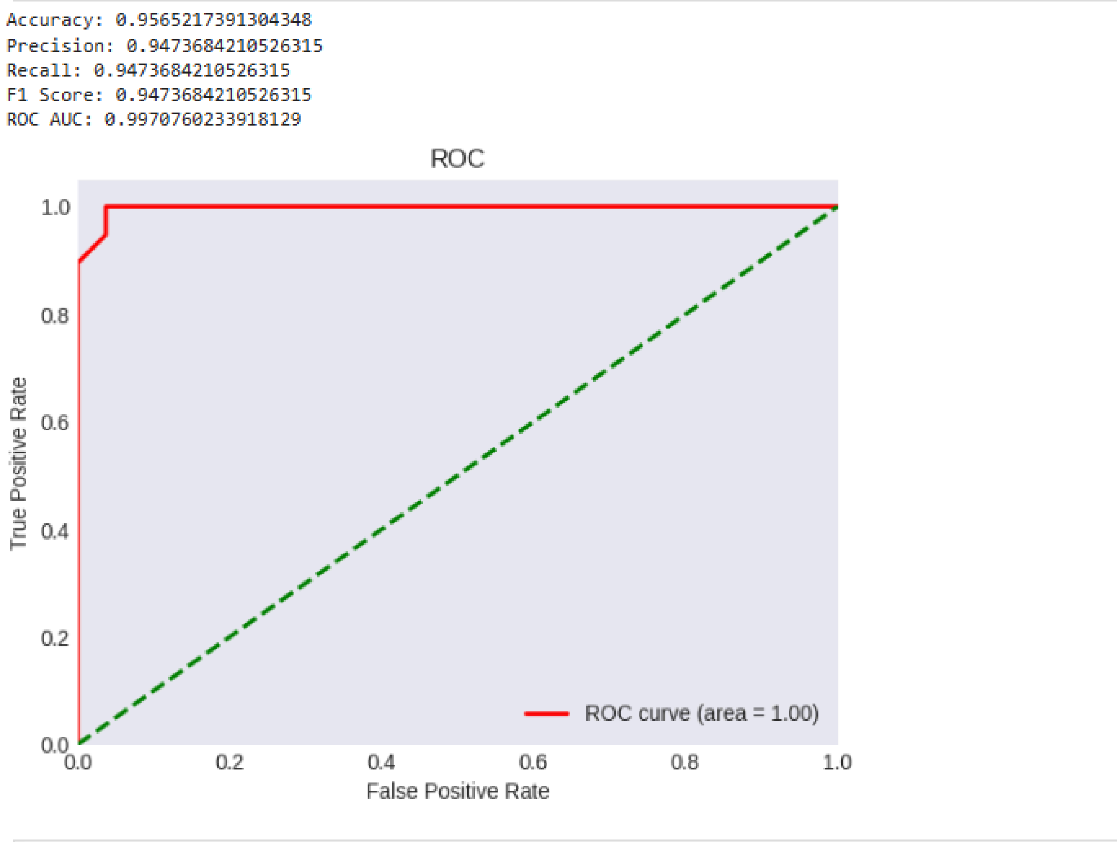


Logestic-Regression

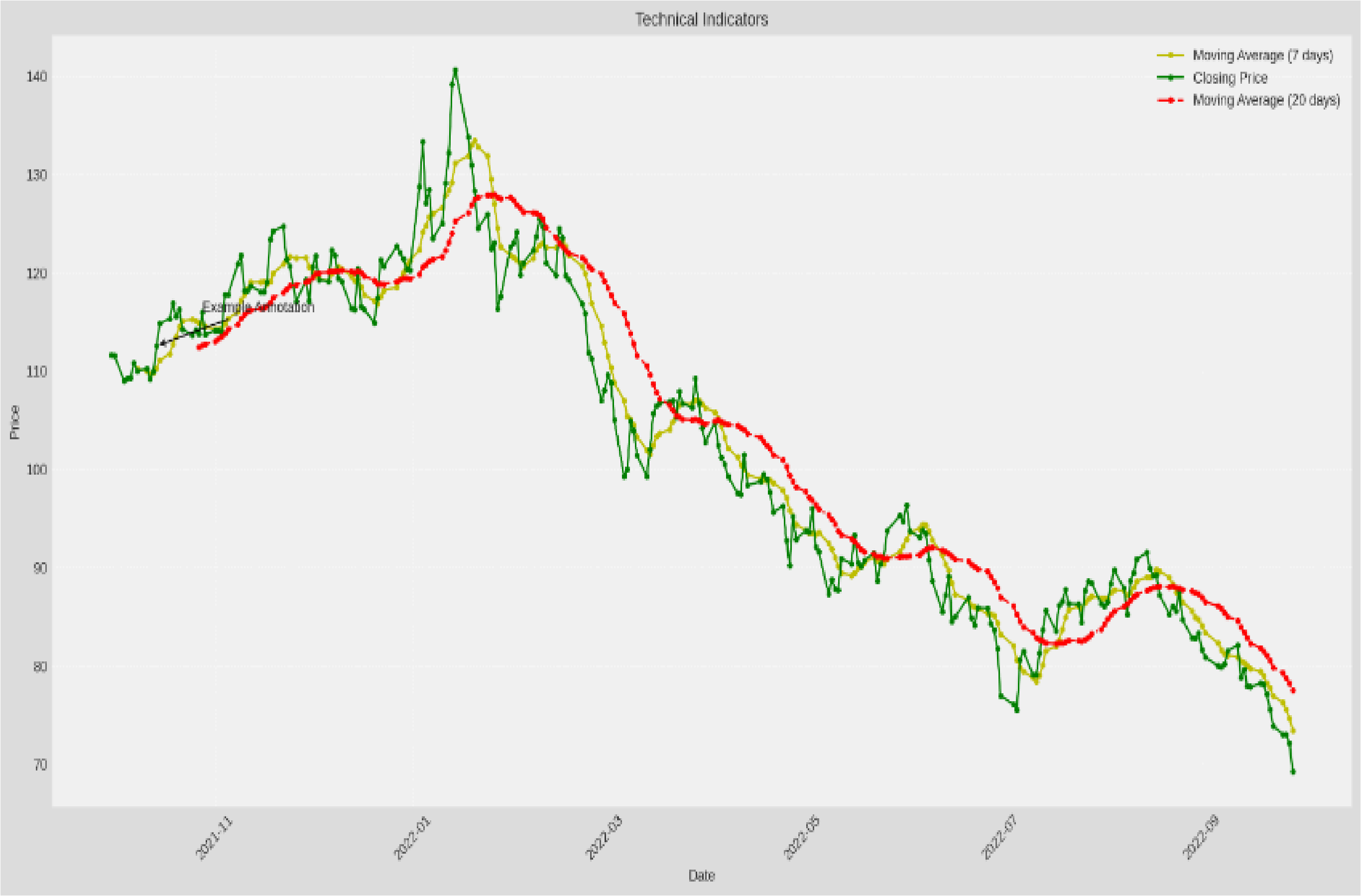


```
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
ROC AUC: 1.0
```

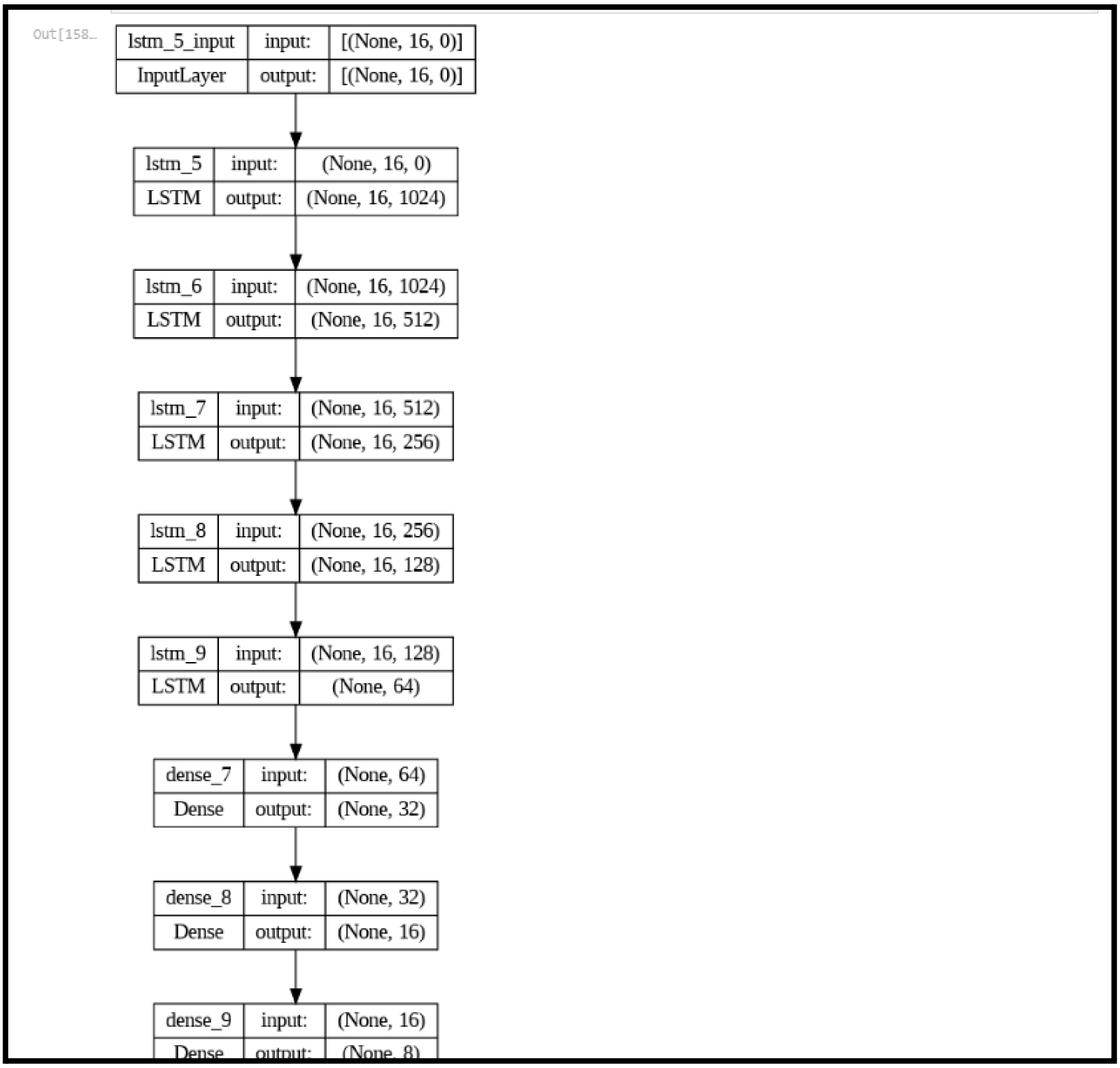
SVM-



```
Accuracy: 0.9565217391304348
Precision: 0.9473684210526315
Recall: 0.9473684210526315
F1 Score: 0.9473684210526315
ROC AUC: 0.9970760233918129
```


Technical Indicators
Moving Average (7 days)
Closing Price
Moving Average (20 days)
Example Annotation
140
130
120
110
100
90
80
70
Price
2021-11
2022-01
2022-03
2022-05
2022-07
2022-09
Date

Keras-modal


Out[158…
lstm_5_input | input: | [(None, 16, 0)]
InputLayer | output: | [(None, 16, 0)]
lstm_5 | input: | (None, 16, 0)
LSTM | output: | (None, 16, 1024)
lstm_6 | input: | (None, 16, 1024)
LSTM | output: | (None, 16, 512)
lstm_7 | input: | (None, 16, 512)
LSTM | output: | (None, 16, 256)
lstm_8 | input: | (None, 16, 256)
LSTM | output: | (None, 16, 128)
lstm_9 | input: | (None, 16, 128)
LSTM | output: | (None, 64)
dense_7 | input: | (None, 64)
Dense | output: | (None, 32)
dense_8 | input: | (None, 32)
Dense | output: | (None, 16)
dense_9 | input: | (None, 16)
Dense | output: | (None, 8)

CONCLUSION

Model Evaluation

The logistic regression model exhibited balanced performance, achieving high recall but lower precision and ROC AUC scores. This indicates that while the model was effective at identifying most positive cases (price increases), it also generated a considerable number of false positives.

To select the best performing model, other models such as SVM should be evaluated similarly. The logistic regression results indicate a need for further tuning or potentially the use of more complex models to enhance precision and overall predictive accuracy.

Portfolio Analysis

The trading strategy based on sentiment analysis produced mixed results. Conversely, the strategy did not perform well for the TSM Inc stock, resulting in a loss of 1.2%. This stark difference in performance underscores the variability in outcomes based on different stocks and suggests that the sentiment-based strategy may need to be tailored or adjusted for specific assets to optimize returns.

RESOURCES

- Algo Trading using Python - Free Code Camp
- https://blog.quantinsti.com/sentiment-analysis-trading/
- S tock Market Sentiment Analysis Using Python & Machine Learnin