REPORT
ON

Stock Market Prediction

Abstract

The stock market is a dynamic and complicated system, it is essential for investors to precisely predict its movements to make wise choices. The stock market is a difficult market to predict, and conventional methods are not always correct or effective. When analysing large amounts of text from financial news stories, Natural Language Processing (NLP) techniques can be used to spot patterns and more precisely forecast stock market trends. This study analyses the text data from financial news articles to predict the stock market trends.

1. Introduction

This report focuses on predicting stock market trends using machine learning algorithms. The study collects and preprocesses data from various sources, including financial news, social media, and other economic indicators. It then extracts relevant features from the data and applies machine learning algorithms to predict stock market trends.

The report discusses the limitations and challenges of stock market prediction and emphasizes the importance of data quality and feature engineering for accurate predictions.

The report evaluates the performance of different machine learning algorithms and compares their accuracy in predicting stock market trends. The study finds that the Linear Discriminant algorithm outperforms other algorithms, such as Linear Regression and Support Vector Machines.

The study concludes that machine learning algorithms can help in predicting stock market trends, but they should be used in conjunction with other economic indicators and expert knowledge to make informed investment decisions.

Overall, the report provides insights into the use of machine learning algorithms in predicting stock market trends and offers recommendations for further research in this area. It highlights the potential of machine learning to aid investors in making informed decisions in a dynamic and complex stock market environment.

2. Methodology

This section describes the step-by-step method used in this project. The aim of this project was to build a sentiment analysis model using news headlines to predict stock market prices. The project involved the following steps:

2.1. Data Gathering:

The first step in building a sentiment analysis model is to gather the necessary data. For this project, two datasets were used. The first dataset contained daily news headlines extracted from Reddit WorldNews Channel (/r/worldnews). They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date.

The second dataset contained the daily stock market prices of the Dow Jones Industrial Average (DJIA). The data was gathered from the Kaggle website.

	Date	Open	High	Low	Close	Volume	Adj Close
0	2016-07-01	17924.240234	18002.380859	17916.910156	17949.369141	82160000	17949.369141
1	2016-06-30	17712.759766	17930.609375	17711.800781	17929.990234	133030000	17929.990234
2	2016-06-29	17456.019531	17704.509766	17456.019531	17694.679688	106380000	17694.679688
3	2016-06-28	17190.509766	17409.720703	17190.509766	17409.720703	112190000	17409.720703
4	2016-06-27	17355.210938	17355.210938	17063.080078	17140.240234	138740000	17140.240234

2.2. Data Cleaning

The next step involved cleaning the news headlines dataset. The headlines had unwanted characters such as slashes, breaks, and quotes. These characters were removed using regular expressions. The headlines were then combined into one column for each day in the dataset. This was done to create a single text block that can be used for sentiment analysis.

2.3 Data Preparation:

The preparation of the data involved several steps. The first step was to calculate the subjectivity of the headlines using the TextBlob library. Subjectivity is a measure of the degree to which the headlines express opinions or emotions rather than facts. This value was calculated for each headline. The second step was to obtain the sentiment scores for each headline using the VaderSentiment library. The VaderSentiment library assigns a value between -1 and 1 to each headline based on its polarity. The compound score represents the overall sentiment of the headline, while the negative, positive, and neutral scores represent the proportion of the headline that is negative, positive, or neutral, respectively.

2.4. Feature Selection

The selection of features is an important step in building a sentiment analysis model. In this project, features such as subjectivity, compound, negative, positive, and neutral sentiment scores were selected for the analysis. Additional features such as the Open price, High, Low and Volume for the day have also been given. These features were selected based on their relevance to the project objective, which is predicting stock market prices based on news headlines and price data.

2.5. Model Building

Several machine learning models were used in this project. The models used were Linear Discriminant Analysis, Logistic Regression and Support Vector Machine. The Linear Discriminant Analysis model was found to provide the best results with an accuracy of 85%. This model is a powerful tool for analyzing large volumes of data and has been widely used in various fields such as finance and medicine.

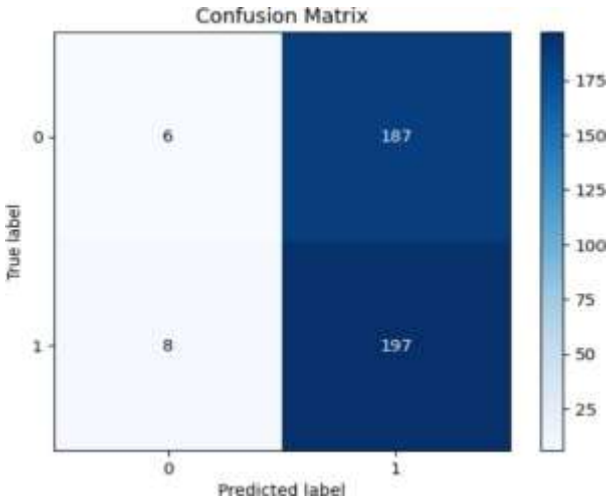
3. Results

The performance of the models was evaluated using the classification_report() function. This function provides a detailed breakdown of precision, recall, f1-score, and support for each class. The confusion matrix was also plotted using the scikitplot.metrics library.

3.1. Support Vector Machine

classification report

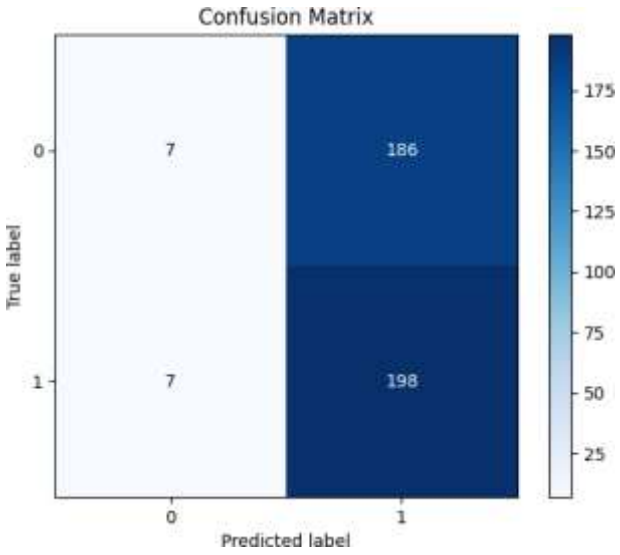
	precision	recall	f1-score	support
0	0.43	0.03	0.06	193
1	0.51	0.96	0.67	205
accuracy			0.51	398
macro avg	0.47	0.50	0.36	398
weighted avg	0.47	0.51	0.37	398



3.2. Logistic Regression

classification report

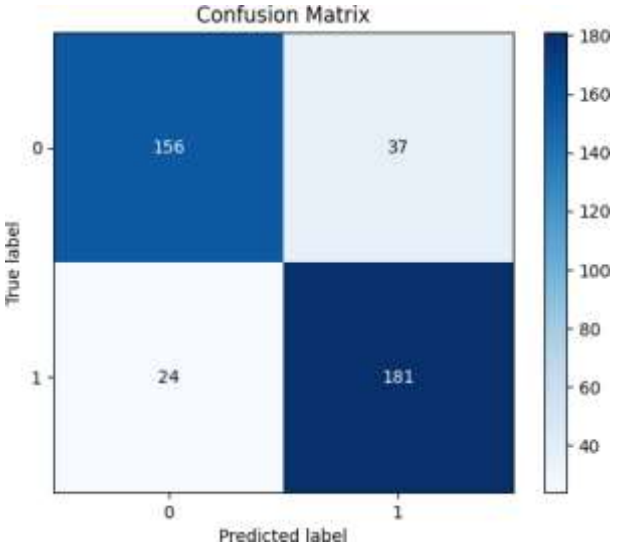
	precision	recall	f1-score	support
0	0.50	0.04	0.07	193
1	0.52	0.97	0.67	205
accuracy			0.52	398
macro avg	0.51	0.50	0.37	398
weighted avg	0.51	0.52	0.38	398



3.3. Linear Discriminant Analysis

classification report

	precision	recall	f1-score	support
0	0.87	0.81	0.84	193
1	0.83	0.88	0.86	205
accuracy			0.85	398
macro avg	0.85	0.85	0.85	398
weighted avg	0.85	0.85	0.85	398



4. Conclusion

In conclusion, the sentiment analysis model built in this project provides a useful tool for predicting stock market prices based on daily news headlines. While the Linear Discriminant Analysis model supplied the best results, there is still room for improvement. Additional features could be added to the model to improve its accuracy. Furthermore, more sophisticated machine learning models could be used to achieve better results. Overall, this project shows the power of sentiment analysis and its potential to revolutionize the prediction of stock market prices.