Spring 2023, SL Term Project

林哲毅 r11922148

Sentiment Analysis of News for Stock Price Movement Prediction

Abstract

Stock market prediction has long been a subject of great interest for researchers and investors alike. While various factors contribute to the movement of stock prices, sentiment analysis of news articles has emerged as a promising avenue for predicting stock price movements. In this research project, we propose a framework that utilizes pretrained models from Hugging Face to predict sentiment scores of news articles. Subsequently, we employ a selection of popular machine learning algorithms, including Support Vector Machines (SVM), logistic regression, probit regression, K-Nearest Neighbors (KNN), decision trees, Naive Bayes, and ensemble methods, to predict the direction of stock price movements. Our study aims to compare the performance of these algorithms and provide insights into the effectiveness of sentiment analysis in stock market prediction.

1. Introduction

The stock market is a dynamic and complex system influenced by a myriad of factors, including economic indicators, company financials, market trends, and investor sentiment. Accurately predicting stock price movements is a challenging task, as it requires identifying patterns and relationships within an ever-changing landscape of information. In recent years, sentiment analysis of news articles has garnered significant attention as a potential tool for stock market prediction. The underlying assumption is that news sentiment can serve as a proxy for investor sentiment, thus impacting stock prices. By leveraging advances in natural language processing (NLP) and machine learning, sentiment analysis models can extract sentiment information from textual data and quantify the positive, negative, or neutral sentiment associated with news articles. This paper presents a study that employs pretrained models on Hugging Face to predict sentiment scores of news articles, followed by the utilization of popular machine learning algorithms for classification. We compare the performance of Support Vector Machines (SVM), logistic regression, probit regression, K-Nearest Neighbors (KNN), decision trees, Naive Bayes, and ensemble methods in predicting the direction of stock price movements based on sentiment scores. Through this investigation, we aim to provide valuable insights into the feasibility and effectiveness of utilizing sentiment analysis in stock market prediction, thereby contributing to the growing body of literature in the field of financial forecasting.

2. Dataset

A. Raw Dataset

The dataset used in this research project comprises both news data and price data. The raw dataset was obtained from Kaggle, a popular platform for sharing and discovering datasets.

- The news data consists of historical news headlines extracted from the Reddit WorldNews Channel (/r/worldnews). These headlines are ranked based on the votes of Reddit users, and only the top 25 headlines for a given date are considered. The news data spans the period from 2008-08-08 to 2016-07-01, providing a substantial timeline for analysis.

- The price data is based on the Dow Jones Industrial Average (DJIA), a widely recognized stock market index that tracks the performance of 30 major publicly traded companies in the United States. The price data aligns with the same timeframe as the news data, enabling a comprehensive analysis of the relationship between news sentiment and stock price movements.

B. Dataset Quickview

Data	Dimension	Columns
Price data	(1989, 7)	• The columns include Date, Open, High, Low, Close, Volumn, Adj Close
News data	(1989, 26)	• The columns include Date, Top1, ..., Top25 • Top1 means the most voted news, Top2 means the second most voted news

Table 1: Dimension and columns of the two dataset

3. Methodology

A. Features and Label

- Features are sentiment scores of news of the previous day. We have 25 news per day, and we will use sentiment models to create sentiment scores.
- Label is the movement direction of the current day. We compare the open price and the adjusted close price. If the open price is bigger, then the label is 0; otherwise, the label is 1.

B. Pre-trained Models for Sentiment Scores

In this research project, pretrained models are utilized to predict sentiment scores of news articles. Pretrained models are pre-trained on large-scale datasets and have learned valuable representations of language, enabling them to perform various NLP tasks with minimal fine-tuning. For this study, we leverage the vast capabilities of

pretrained models available on the Hugging Face website, a prominent platform for accessing and sharing pretrained models and NLP resources.

I utilized a single pretrained model, specifically cardiffnlp/twitter-roberta-base-sentiment, for sentiment score prediction. This model has been trained on a diverse dataset, including social media data, making it suitable for capturing sentiment information from news articles in the context of stock market prediction.

To obtain sentiment scores, the news articles are inputted into the pretrained models. The first two models will generate probabilities for positive, neutral, and negative sentiments. The probabilities indicate the likelihood of each sentiment being present in the respective news article. We adopt a scoring mechanism where the sentiment score of a news article is determined by subtracting the probability of a negative sentiment from the probability of a positive sentiment. This scoring mechanism ensures that the sentiment scores range from -1 (indicating highly negative sentiment) to 1 (indicating highly positive sentiment).

Label	Probability
positive	0.8466
neutral	0.1458
negative	0.0076

Table 2: Example output of twitter-roberta-base-sentiment model for input “Good night”

index	label	top1_score	top2_score	…	top24_score	top25_score
1						
2						
…						

Table 3: Overview of the label and calculated features. For example, top1_score means the calculated score of top1 news based on the output of the pretrained model

C. Feature Selection

Feature selection is a crucial step in machine learning, aimed at identifying the most informative and relevant features. In my research, I employed Ridge regression to select the top 10 sentiment scores. Ridge regression helps shrink the coefficients of less important features towards zero, allowing me to focus on the most influential factors affecting the target variable. This approach improves model accuracy and efficiency while providing valuable insights into the key features that drive predictions.

D. Classification Methods

In this research project, a classification approach is employed to predict the direction of stock price movements based on the sentiment scores of news articles. The dataset is divided into training and testing sets, with 82% of the data allocated for training and the remaining 18% for testing.

Six popular machine learning algorithms are chosen for classification: Support Vector Machines (SVM), logistic regression, probit regression, Naive Bayes, K-Nearest

Neighbors (KNN), and decision tree. These algorithms offer distinct methodologies for pattern recognition and classification tasks.

Training / Testing Splig	Models
• 82% of data for training = 1589 observations • 18% of data for testing = 348 observations	SVM
	logistic regression
	probit regression
	Naive Bayes
	KNN
	Decision Tree
	Ensemble Approachs

Table 4: Settings of the experiment

To further enhance the classification performance, this study employs ensemble approaches by combining the predictions of the four individual models using three different methods: hard voting, soft voting, and stacking. These ensemble methods aim to leverage the collective decision-making of the individual models and improve the accuracy and robustness of stock price movement predictions based on sentiment scores derived from news articles.

In the hard voting ensemble approach, the final prediction is determined by majority vote among the individual models. Each model independently predicts the class label, and the class with the most frequent prediction across the models is selected as the final prediction.

The soft voting ensemble approach takes into account the probabilities assigned by each individual model. The predicted probabilities of the models are averaged, and the class with the highest average probability is chosen as the final prediction. This approach considers the confidence levels of the models in addition to their predictions.

The stacking ensemble method involves training a meta-model that learns from the predictions of the individual models. The predictions of the models serve as additional features for the meta-model, which then makes the final prediction. This approach aims to capture the strengths of the individual models by combining their predictions in a higher-level model.

E. Methods that did not work well

Three specific approaches that were attempted but did not perform well are discussed below. All the three approaches were trying to add more features.

The first method involved incorporating multiple days of news sentiment as features. The rationale behind this approach was to capture a broader context and potential cumulative effect of sentiment over consecutive days. However, despite the initial intuition, this method did not lead to improved prediction accuracy.

Another approach that was explored was the inclusion of 83 technical indicators as features. Technical indicators are widely used in financial analysis to identify patterns and trends in price data. However, despite the extensive set of indicators used,

incorporating them as features did not lead to significant improvements in prediction performance.

The last method was to include sentiment features from more pretrained models. The additional sentiment features did not contribute significantly to capturing the price movements, resulting in limited predictive power. So we chose to use only one pretrained model for sentiment score in the final setting.

While these methods did not yield the desired outcomes in terms of prediction accuracy, their exploration serves as valuable insights for understanding the limitations and challenges associated with sentiment analysis in the context of stock price movement prediction. These findings emphasize the importance of carefully selecting relevant features and considering the specific characteristics of the data and domain to enhance the effectiveness of sentiment analysis approaches in financial forecasting.

4. Results

Among these 6 models, logistic regression and probit regression demonstrate the highest accuracy in predicting the direction of stock price movements on the testing set based on the sentiment scores derived from news articles.

Model	Training Accuracy (%)	Tesing Accuracy (%)
SVM	67.02	58.91
Logistic Regression	55.07	62.07
Probit Regression	55..07	61.78
Naive Bayes	55.63	60.63
KNN	74.01	56.90
Decision Tree	59.16	56.32
Ensemble Hard Voting	65.70	63.22
Ensemble Soft Voting	71.56	61.49
Ensemble Stacking	74.83	59.20

Table 5: Training accuracy and testing accuracy of the 6 models and the ensemble methods

It is noteworthy that all six individual models surpass the baseline accuracy of 50%, which represents the accuracy of random guessing. This suggests that the models possess some degree of predictive power in capturing the relationship between sentiment scores and stock price movements.

To further enhance the classification performance of predicting stock price movements based on sentiment scores, three ensemble methods were employed: hard voting, soft voting, and stacking. Among these ensemble methods, the hard voting approach achieved the highest accuracy of approximately 63%.

The results of this research project demonstrate the potential efficacy of using sentiment analysis of news articles for predicting stock price movements. The ensemble approach, in particular, showcases its superiority in terms of accuracy, highlighting the benefits of aggregating predictions from multiple models.

5. Conclusion

In conclusion, this research project explores the challenging task of predicting stock price movements based on news sentiment. This hard-voting ensemble method achieves an accuracy of 63.22%. This underscores the power of ensemble techniques in improving classification performance and minimizing overfitting risks.

Nevertheless, it is crucial to acknowledge the complexities and uncertainties inherent in stock price prediction, particularly when relying solely on news sentiment. While the obtained results show promise, further advancements can be made by incorporating additional factors such as economic indicators, market trends, and company-specific information. Additionally, future research endeavors could explore the utilization of more extensive datasets and more advanced models to further enhance predictive capabilities in this domain.

Reference

1. Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved May, 2023 from https://www.kaggle.com/aaron7sun/stocknews.
2. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification (Barbieri et al., Findings 2020)
3. cardiffnlp (2023). twitter-roberta-base-sentiment. Retrieved from https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
4. cardiffnlp (2023). twitter-roberta-base-sentiment-lates. Retrieved from https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
5. yiyanghkust (2022). finbert-tone. Retrieved from https://huggingface.co/yiyanghkust/finbert-tone
6. OpenAI. (2022). GPT-3.5 Model. Retrieved from https://www.openai.com/models/gpt-3.5/
7. hktseng. (2023). Statistical Learning. Retrieved from https://rpubs.com/hktseng