Prediction of the stock price using Machine Learning techniques

Apoorva Y.
apoorvayelburgi123@gmail.com
Rajarajeswari College of Engineering, Bengaluru, Karnataka

ABSTRACT

As people's interest in forecasting stock prices has been increased in recent years, research on stock price analysis using big data and artificial intelligence. In this paper, we performed sentimental analysis by this work by creating and analyzing a sentimental vocabulary using news items. We can get the positive index of news stories using the sentimental dictionary. We can get the positive index of news stories for each date using the emotive dictionary. We can get the positive index of news stories for each date using the emotive dictionary. We can confirm the utility and possibility of sentimental analysis in the stock market by examining the correlation value between the positive index value and the stock return value.

Keywords—*Stock Price Analysis; Sentimental Analysis; Sentimental Dictionary; Positive Index Value; Stock Return Value Are Some Of The Terms Used In This Paper.*

1. INTRODUCTION

Machine learning techniques are now being researched in a variety of fields. Stock investors may profit if they correctly estimate stock price movement. However, constructing good predictive models is incredibly difficult due to the intricacies of stock market data. We also have the option of choosing from a variety of data sets for the study. This is due to the large number of available options.

In addition to machine learning techniques, there is a sentimental analysis method that employs Natural Language Processing (NLP). Create models for categorising "tweets" into good, negative, and neutral attitudes, evaluating product opinions expressed in blog posts, comments, and reviews, and so on. As a result, sentimental analysis has become increasingly popular. Natural language processing (NLP) has the potential to be extremely beneficial in a variety of situations. Text data, such as news stories, would have a significant impact on the stock market.

Stock prices are likely to be considerably impacted. We can establish whether the data had an impact on the stock price in this study using emotional analysis. There are several approaches to sentimental analysis, one of which is Word2Vec.

In addition to this method, creating a sentimental lexicon is another way to do sentimental analysis. There are two approaches. The first is to use KOSAC (Korean Sentiment Analysis Corpussentimental vocabulary, and the second is to generate a new dictionary from news articles. In this work, we propose that we develop a completely new dictionary. This is the case because the words usually used in stock news items differ from the words in the KOSAC dictionary.

The data for the study came from the website 'bigkinds' (https://www.bigkinds.or.kr/). Among the countless equities, we gathered news pieces from bio-related companies. This is because we expected news stories to have a big impact on biotech stocks. The sentimental dictionary was created using two years of news stories.The test data consists of news articles regarding bio regenerative medicine. Stocks in the last three months.We use a Korean morpheme analyzer called KKMA because our data is text data in Korean (Kind Korean Morpheme Analyzer).

2. DESIGN AND IMPLEMENTATION

Construction of sentiment dictionary
There are three methods for creating a sentiment dictionary. Deep learning approaches include word hand-tagging, positive index algorithms, and deep learning techniques like Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). The second strategy, which is extremely simple and clear, is presented in this paper. Newspaper articles and stock prices for each date are necessary for the study. We employ Day stock prediction (NSP) and stock price return to calculate stock prices (SPR).
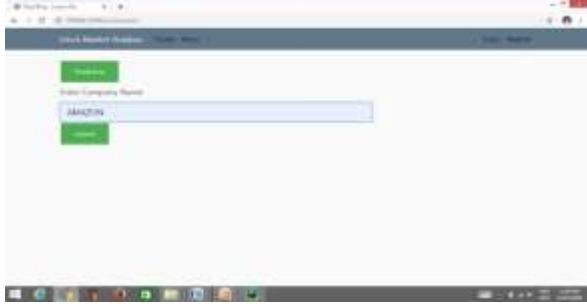
NSP is a metric that shows if stock prices are expected to rise in the coming months. The SPR is the difference in price between the previous day's closing price and today's closing price.

Python
Python is a widely used interpreted high-level, general-purpose dynamic programming language. Its design philosophy emphasises code readability, and its syntax allows programmers to express concepts in fewer lines of code than languages like C or Java. The language offers elements that allow simple and large-scale programmes to be written in a succinct manner. As a

result, the programming language is suitable for both desktop and web applications. Python can also be used to build complex scientific and numerical programmes. Python includes tools to help with data analysis and visualisation.

NLTK (Natural Language Tool Kit)

For working with human language data, NLTK is a popular Python programming language. It comes with a set of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, as well as wrappers for industrial-strength NLP libraries, an active discussion forum, and easy-to-use interfaces to over 50 corpora and lexical resources like WordNet. NLTK has been described as "a wonderful tool for teaching and working in computational linguistics using Python," as well as "an amazing library to explore with natural language."

3. MODELS
1. Register/Login
2. Prediction

Register/Login
This programme will require users to register with their user name, email address, and password. After you've successfully finished the signup process. After finishing the registration process, the user can log in to the system. The user can upload the names of companies such as IT firms or marketing firms after logging in with his or her e-mail address and password.The company name is displayed to other users after it is uploaded.

Prediction
Users who want to forecast the fate of their company should first upload the company's name and then click the predict button. The stock price prediction will be sent to the user. After the operation is finished, the user can logout of this programme.

4. SNAPSHOTS

Fig: Login page

The above image depicts the login page, which requires users to input their e-mail address and password. The user is authenticated by connecting to the database and checking whether or not he is an authorized user.

Fig: Register page

The above image depicts the registration page, where users will enter their user name, email address, and password.   Fig: prediction page

The top module demonstrates where users can enter company names such as IT firms or marketing firms. The names of the companies are displayed to the other users when they have been uploaded.
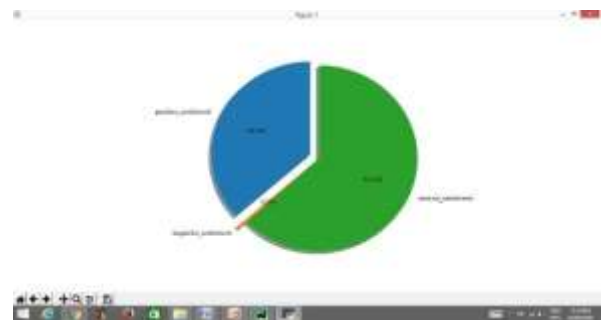

Fig: Result page

By entering his company name on the above website, the user will receive a forecasted outcome in the form of a pie-chart.

Fig: Result page

5. EXPERIMENT RESULTS
The words in the sentiment dictionary have a positive index value (PI). On the basis of frequency, the words were sorted in descending order. As a result, we can use the dictionary to see which words appear frequently in news articles. The top ten most often used words in the dictionary are listed in Table 1. Common sense can also reveal that those are the terms we'll see a lot of when we read news article about stocks.

The creation of a sentiment dictionary is now complete. This lexicon was used to create fresh news stories. First, assess the morphology of the data, and then extract the nouns. Calculate the positive index of the text by comparing the extracted nouns to the terms in the sentiment dictionary. The positive index of text (PT) is the average value of the positive index of extractive nouns (PT).is the positive index of text (PT) is the average value of the nouns taken from the text's positive index.

Table I: Top 10 Words in the dictionary

Word	Positive index
일	0.4470
코스	0.4506
피	0.4488
지수	0.4479
시	0.4569
포인 트	0.4470
바이 오	0.4410
코스 피	0.4515
코스 닥	0.4410
원	0.4510

When a dictionary has too many terms, it may include words that are rarely used. It may be impossible to calculate the exact positive index. As a result, several tests were conducted as part of this project. In a dictionary, increasing or decreasing the number of terms. The results are shown in the graph below.
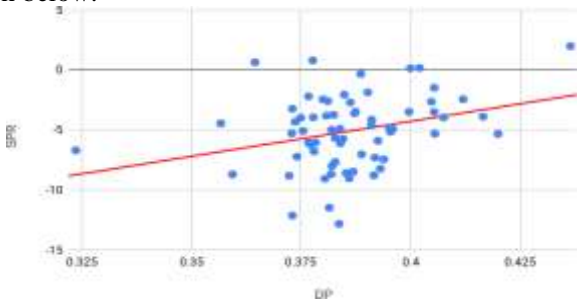

Fig: The correlation value between the positive index (PI) and the next day stock return value is shown in Figure. (SPR)

As can be observed in Fig, the correlation value has increased slightly when there are 8000 words in the dictionary compared to 10000 words in the dictionary. However, we can confirm that the correlation value decreases as the number of words decreases. We can see how difficult it is to accurately predict stock price movements due to the low frequency of some of the phrases. Some words may have a high positive value due to their rarity. In addition, if the dictionary has too few terms, the chances of finding the word used in the news story in the dictionary are reduced. The correlation value drops as a result. As a result, the best scenario was determined to be the 8000 words listed in the dictionary.
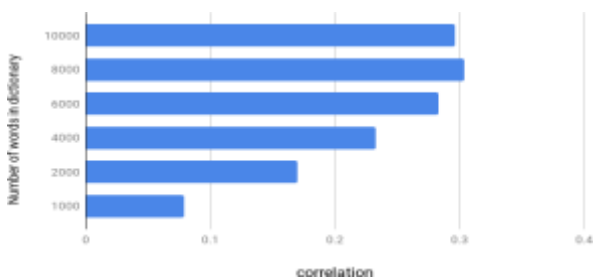

Fig: DP is on the x-axis, and SPR is on the y-axis. The correlation coefficient is approximately 0.3034.

Fig illustrates a graph of the 8,000 terms in the sentiment lexicon with the highest correlation value. The daily positive index (DP) for each day is shown on the x-axis, while the value of the next day's stock price return is shown on the y-axis (SPR). Approximately 0.3034 is the correlation coefficient. This may be observed in the red trend as well.

6. CONCLUSION

In this work, we use news articles to predict stock price using emotional analysis. The emotional analysis strategy we utilised was to create a new sentiment dictionary based on the news stories. Through this, we obtained the correlation between the positive index for each date (DP) and the return value of the next day's stock price (SPR). This can be used to determine the trend of stock price fluctuations. Based on the experiment, a correlation value of 0.3034 was calculated. Because of the nature of the Korean language, deducing its high accuracy using natural language processing methods rather than English is more difficult. As a result, this level of correlation in the extremely complex stock data is seen as quite important. If we remove additional superfluous phrases from the code and perform a range of trials with different sentiment dictionaries, we can build a more efficient sentiment dictionary.

7. REFERENCES

[1] University of California, San Diego, G. Gidófalvi and C. Elkan, "Using News Articles to Predict Stock Price Movements," June 15, 2001.

[2] Y. Kara, M. A. Boyacioglu, and K. Baykan, "Using artificial neural networks and support vector machines to predict the direction of stock price index movement: The sample of the Istanbul Stock Exchange," Expert Systems with Applications, vol. 38(5), May1, 2011.

[3] K. Kim, "Financial time series forecasting with support vector machines," Neurocomputing 2003, DOI: 10.1016/S0925-2312(03)00372-2.

[4] "Using Trend Deterministic Data Preparation and Machine Learning Techniques to Predict Stock and Stock Price Index Movement," Expert Systems with Applications: An International Journal, vol. 42, no. 1, January 2015. J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Using Trend Deterministic Data Preparation and Machine Learning Techniques to Predict Stock and Stock Price Index Movement," Expert Systems with Applications: An International Journal, vol. 42, no. 1, January 2015, pp. 259-268.

[5] Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data," LSM '11 Proceedings of the Workshop on Languages in Social Media, pp. 30-38

[6] G. Vinodhini and R.M. Chandrasekaran, " Sentiment Analysis and Opinion Mining: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 6, June 2012.Z. Su, H. Xu, D. Zhang, and Y. Xu, "Chinese emotion categorization using a neural network tool – Word2Vec," Expert Systems With Applications, vol. 42, issue 4, pp. 1857-1863.

[7] D. Kim, J. Park, and J. Choi, "A Comparative Study of Stock Price Prediction Models Based on Social Media and News Articles Using Sentiment Analysis and Machine Learning," Korea IT Service Association, vol.12, no. 3 [2014], pp. 211-233.

[8] M. Cliché, "BB twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs," Proceedings of SemEval-2017, April 20, 2017.