STOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS

Abstract

The Efficient Market Hypothesis is an important concept of financial economics that states that share prices reflect all the information and are used for stock prediction. Financial news articles play an important role in investor sentiments and impact the decision made by investors and therefore change the market state. There are infinite amounts of news sources in today’s digital age. It's humanly impossible to read and find all relevant information in the form of news to draw a conclusion timely to make an investment plan that returns maximum profit. In this project, we are taking non-quantifiable data such as financial articles about listed stocks or companies, and predicting their future stock trend with news sentiment classification. We are using different classification models to predict whether the news will have a neutral, positive, or negative impact on the stock price.

1. INTRODUCTION

In the finance field, the stock market and its trend are highly volatile in nature. Stock returns display time-varying serial correlation which attracts researchers to capture the volatility and predict its next moves. Many studies have analyzed third issues using a wider range of variables and techniques. In spite of growing research, empirical evidence suggests that fundamental factors of time-varying risk trading are not sufficiently large to explain autocorrelation observed in stock returns. Mainly there are two methods for forecasting market Technical and Fundamental analysis. The technical analysis considers past prices and volume to predict future movement. Fundamental analysis of a business involves analyzing its financial data to get some insights.

The motivation of this paper is to build and check the impact of news articles on stock prices. We are using different Natural Language Processing models(text mining) and supervised machine learning as a classification to check the news polarity. And also be able to classify unknown news, which is not used to build a classifier. Different classification models are implemented to check and improve classification accuracy. We have datasets that contain the sentiments for financial news headlines from the perspective of a retail investor.

This Paper is arranged as follows. Section 2 provides an overview of the literature concerning the prediction of the stock market, textual representation and sentiment analysis techniques. Section 3 contains the method of Preprocessing of data and Sentiments Detection Algorithm that provides all the insights information from news articles in the dataset. Section 4 details our experimental findings and discusses their impact on stock market prediction. Section 5 delivers our conclusions and a brief discussion of future research directions.

2. LITERATURE SURVEY

Over the past many years, Important changes have taken place in the environment of the financial market. The stock price trend is an active research area, as more accurate predictions are directly related to more returns in stocks. In recent years, significant efforts have been put into developing models that can predict the future trend of a specific stock market. A growing number of research papers use the NLP method to assess how sentiments of firm-specific news, financial reports, or social media impact stock market returns. Some researchers showed that there is a strong relationship between news articles about a company and its stock price fluctuations.

An important early work (2007) by Tetlock[4] explores the possible correlation between the media and the stock market using information from the Wall Street Journal and finds that high pessimism causes downward pressure on stock prices.

Nagar and Hashler in the research[5] presented an automated text mining-based approach to aggregate news from various sources and create a news corpus. The Corpus is filtered down to relevant sentences and analyzed using NLP techniques. NewsSentiment utilizing the count of positive and negative polarity words is proposed as a sentiment measure of the overall news corpus. They have used various open-source packages and tools to develop the news collection and aggregation engine as well as the sentiment evaluation engine. They also state that the time variation of NewSentiment shows a very strong correlation with the actual stock price trend.

Kogen et al.[6] use SVM to predict the volatility of stock market returns. The results indicate that text regression algorithm prediction correlates with true volatility nearly as well as historically, and combined algorithm to perform better.

Mahajan et al.[7] used Linear Discriminant Analysis(LDA) to identify topics of financial news and then to predict an up or down in the stock market based on topics extracted from financial news.

Yu et al.[8] present a text mining-based framework to predict the sentiments of news articles and demonstrate its impact on energy demand. News sentiments are quantified and then presented as a time series and compared with fluctuations in energy demands and prices.

Calomiris et al.[9] use news articles to develop a methodology to predict risk and return in the stock market in developed and emerging countries. Their results indicate that the topic-specific sentiments, frequency and unusual news text can predict future returns and drawdowns.

J.Bean[10] uses keyword tagging on Twitter feeds about airline satisfaction to score them for polarity and sentiment. This can provide a quick idea of the sentiment prevailing about airlines and their customer satisfaction ratings.

Jiao et al.[7] show that high social media activity around a specific company predicts a significant increase in return volatility whereas attention from influential press outlets e.g. the Wall Street Journal in fact is a predictor of the opposite: a decrease in return volatility.

3. METHODOLOGY

This section gives details about the design of the news based predictive system. It explains how news categories were specified, describes raw textual data, and discusses the data pre-processing and the performance of different classification models used for evaluation.

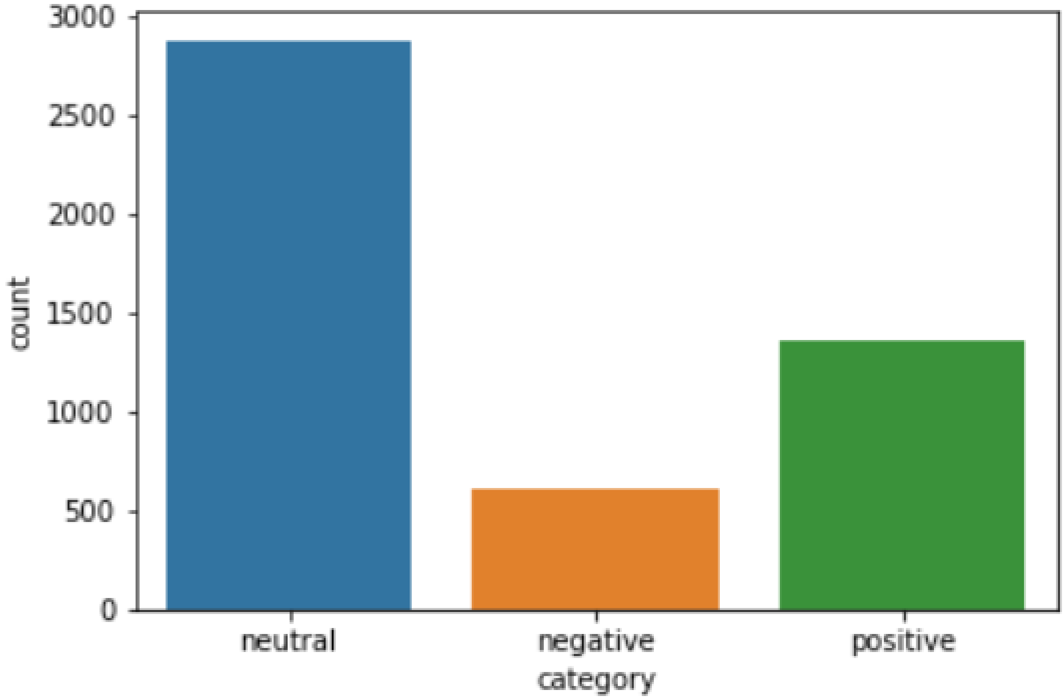
3.1 Data

The data is taken from Kaggle and it contains the sentiments for the financial news headlines from the perspective of a retail investor.

3.2 Sentimental Analysis of Financial News

First we import all the python libraries required like pandas, numpy, seaborn, matplotlib, nltk, string, re and countvectorizer from sklearn library. The next step is to import the dataset that is taken from kaggle and perform some exploratory data analysis which includes seeing the number of rows and columns in the dataset, the data type of entries in the columns and number of negative, positive and neutral news.

The following bar chart shows the number of positive, negative and neutral category of news statements in the dataset:



3.3 Text Preprocessing of Financial News

Text data is unstructured data. So we cannot provide raw text data to the classifier as an input. Firstly, we need to remove the punctuations from the text data, extra spaces and digits since these are unnecessary for classification of text data. After this we remove the stopwords in english such as in, out, before, after, during, do, does etc. using nltk library as these are not useful for the stock price prediction. Then we perform stemming and lemmatization of text data which are necessary to normalize text and prepare words and documents for further processing. For example, like and liked are the same words in different tenses.

After this we perform count vectorization of text data which includes breaking down a sentence into words by performing preprocessing tasks like converting all words to lowercase and removing special characters. Count vectorization makes it easy for the text data to be directly used by the machine learning models for text classification. After this we remove the most frequent words in the sentences using re or regular expression library in python. After this we perform label encoding of categorical variables so that machine learning algorithms perform better in terms of accuracy and other performance metrics when data in numeric form rather than categorical form. After performing all the above steps we can now apply the various machine learning models for text classification and prediction.

3.4 Classification Models

Our data is classified into three different categories which is positive, negative and neutral where positive mean the news will have positive impact (i.e. Price will increase) on stock price, negative mean the news will have negative impact(i.e. Price will decrease) and neutral show that price will remain same or it will not be affected by the news. We used different classification models like Support Vector Machine, Logistic Regression, Multinomial Naive Bayes, Bernoulli Naive Bayes, Gradient Boosting, XGboost, Decision Tree and K-nearest neighbour. The data which is news headings in our case basically classify into three different sentiments positive, negative and neutral. We compared the classification models with our NLP sentiment results and checked the accuracy. Among the all above classification models we found out that SVM model predicts the text classification at a better rate of accuracy.

3.5 System Evaluation

For the evaluation we divided the data into train and test data. We evaluate different classification models performance by checking each one's accuracy, precision, recall and f1-score. The results are given in the next section.

4. EVALUATION

We evaluated the different classification models. In the below table the average is different categories (positive, negative and neutral) average with precision, recall and f1-score.

We found out that the Linear Support Vector Machine model gave us the best accuracy with accuracy 74.14% while Gradient Boosting gave the lowest accuracy (60.45%) compared to other models.

	Decision Tree model with accuracy 70.29%		
	precision	**recall**	**f1-score**
Macro avg	0.72	0.52	0.56
Weighted avg	0.71	0.70	0.66

	Logistic Regression model with accuracy 73.66%		
	precision	**recall**	**f1-score**
Macro avg	0.75	0.58	0.62
Weighted avg	0.74	0.74	0.71

	Linear SVC model with accuracy 74.14%		
	precision	**recall**	**f1-score**
Macro avg	0.70	0.65	0.67
Weighted avg	0.73	0.74	0.73

	K Nearest Neighbour model with accuracy 70.01%		
	precision	**recall**	**f1-score**
Macro avg	0.64	0.58	0.60
Weighted avg	0.68	0.70	0.69

	XGBOOST model with accuracy 69.53%		
	precision	**recall**	**f1-score**
Macro avg	0.68	0.52	0.54
Weighted avg	0.69	0.70	0.65

	Multinomial Naive Bayes model with accuracy 69.26%		
	precision	**recall**	**f1-score**
Macro avg	0.76	0.47	0.47
Weighted avg	0.72	0.69	0.63

	Bernoulli Naive Bayes model with accuracy 70.77%		
	precision	**recall**	**f1-score**
Macro avg	0.76	0.50	0.50
Weighted avg	0.72	0.71	0.66

	Gradient boosting model with accuracy 60.45%		
	precision	**recall**	**f1-score**
Macro avg	0.20	0.33	0.25
Weighted avg	0.37	0.60	0.46

5. CONCLUSION

Our study explores whether the different news categories can provide an advantage in financial decisions based on news sentiments. Sentimental analysis plays an important role in finance and economic prediction. News articles are the best tool to capture the sentiment about the current market, from the automation of sentiment detection and based on the words in the news articles, we can get an overall polarity of news article headlines. If the news is positive, then we can conclude that this news is good for the market, which implies stock prices will go high and if the news is negative, then it may impact the stock price to decrease and if the news is neutral, then there is no impact on the stock price.

From the evaluation of all machine learning models we can infer that the linear SVC model predicts the text classification at a better accuracy rate than other models.

REFERENCES

[1] Yauheniya Shynkevich, T.M. McGinnity, Sonya Coleman, Ammar Belatreche, Predicting Stock Price Movements Based on Different Categories of News Articles, 2015 IEEE Symposium Series on Computational Intelligence

[2] P. Hofmarcher, S. Theussl, and K. Hornik, Do Media Sentiments Reflect Economic Indices? Chinese Business Review. 2011, 10(7): 487-492

[3]Abbasi, A., H. Chen and A. Salem 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems 26(3).

[4] Paul Tetlock. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance 62 (2007). https://doi.org/10. 2139/ssrn.685145

[5] Anurag Nagar, Michael Hahsler, Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams, IPCSIT vol. XX (2012) IACSIT Press, Singapore

[6] Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting Risk from Financial Reports with Regression. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Boulder, Colorado) (NAACL ’09). Association for Computational Linguistics, USA, 272–280.

[7] Anuj Mahajan, Lipika Dey, and S. K. Mirajul Haque. 2008. Mining Financial News for Major Events and Their Impacts on the Market. In 2008 IEEE / WIC / ACM International Conference on Web Intelligence, WI 2008, 9-12 December 2008, Sydney, NSW, Australia, Main Conference Proceedings. IEEE Computer Society, 423–426. https://doi.org/10.1109/WIIAT.2008.309

[8] W.B. Yu, B.R. Lea, and B. Guruswamy, A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting, International Journal of Electronic Business Management. 2011, 5(3): 211-224

[9] Charles W. Calomiris and Harry Mamaysky. 2018. How News and Its Context Drive Risk and Returns around the World. Technical Report. Columbia Business School. https://doi.org/10.2139/ssrn.2944826

[10] J. Bean, R by example: Mining Twitter for consumer attitudes towards airlines, In Boston Predictive Analytics Meetup Presentation, 2011