What is the Relationship between News and Stock Price Trend?

Tianxin Zheng
June 2019

Abstract: With the advancement of machine learning and NLP (Natural Language Processing) techniques, there are more and more new researches focusing on the analysis of the news content on stock volatilities and returns. In this paper, the relationship between news and stock price trend will be investigated using natural language processing techniques. Different word representations and machine learning methods will be compared to evaluate the effect of past news text data on stock market movement, in order to make predictions for future stock trends.

Keywords: News, Stock Price, Natural Language Processing, Machine Learning

1 Introduction

The relationship between news and financial markets has long been a popular research topic. There are many empirical studies investigating into the relationship between news and stock market returns and volatilities. Recently, with the development of machine learning and natural language processing techniques, more and more researches focus on extracting information from news text data to make predictions on stock market movements. This paper will also utilize different machine learning models and natural language processing techniques to predict the stock market trends from news text data. Specifically, the paper will focus on the comparison of different word representation techniques and investigate into the prediction power of different approaches.

2 Literature Review

2.1 News and Financial Markets

There are different theories about predicting the future price of stock market. Fama's Efficient Market Hypothesis (EMH) states that asset price fully reflect all available information, thus it is impossible to continuously outperform the market (Fama, 1964). In Random walk theory, stock market price is also believed to be impossible to predict as the stock price is determined randomly and followed a random walk (Malkiel, 1973).

While these theories state that it is impossible to predict stock market movements, there are different trading philosophies emerging and many of them have developed techniques to predict stock prices from financial news. As early as in 1988, Cutler et

al.(1988) have investigated into the impact of different kinds of news on the variance of stock returns. In 1992, Campbell and Hentschel (1992) proposed an asymmetric model which measures the feedback effect of changing variance on stock returns, further stating that good news would increase stock prices and bad news would lower stock prices. There are also some researches analyzing real stock data and news article and stating that there is a brief time period before the market correct to itself to equilibrium within which there exits a weak ability to predict for future prices (LeBaron et. al, 1999; Gidofalvi, 2001).

More recently, there are more empirical studies conducted on the relationship between news and financial markets. In 2011, Engelberg et al. (2011) analyzed the causal impact of media in Financial Markets by analyzing "behaviors of investors with access to different media coverage" (Engelberg et al., 2011, pp 4). After controlling for other variables, they proved that media coverage could strongly predict trading locally. Similarly, Birz et al. (2011) chose newspaper stories as their measure of news to investigate the impact of macroeconomics news on stock returns. Specifically, they found that news about GDP and unemployment could influence stock returns. Alanyali et al. (2013) used a corpus of daily news from Financial Times from 2007 to 2012 to quantify the relationship between the news mentioned about a particular company and this company's daily transaction volume.

2.2 Natural Language based Financial Forecasting

Recently, with the advancement of machine learning and NLP (Natural Language Processing) techniques. There are more new researches focusing on the analysis of the news content on stock volatilities and returns.

Some sentiment analysis were conducted based on already processed sentiment data provided by some content vendors. Uhl (2014) has investigated into the relationship

between Thomson Reuters Datastream and stock returns and proved there is correlation between them.

Many other researchers conducted their analysis based on some on other techniques. There are different approaches to transform financial text data into features that can be processed by computers for the later modeling phase. One basic and broadly used approach is bag-of-words, which describes the occurrence of each word within a document, thus considering each word count as a feature. Because of its simple nature, there are many researches applying this technique to conduct financial text analysis. For example, Yoshihara et al. (2016) used bag-of-words approach to represent news text and then feed it into a recurrent neural network and RBM model to investigate into the temporal properties of news events for stock market prediction.

To take the context into consideration, some researchers used word embeddings to transform text data to vectors. Word embedding utilizes a dense distributed representation for each word, where words have similar meaning are located closely in vector space. Ding et al. (2015) utilized an event-embedding approach to extract events from news text data and then used a deep convolutional neural network to build a model that could both predict short-term and long-term influence of events on stock market trends.

There are also researches comparing different word representations in textual analysis for prediction. Schumaker et al. (2009) built a text analysis system using bag of words, Noun Phrases and Named Entities as different word representations to conduct news analysis and predict for the discrete stock price. Their results showed that the model with both article terms and stock price at the time the article has been released had the best performance in future stock price prediction. They also compared different word representations and found that a “Proper Noun scheme” performed best among all the different representations (Schumaker et al., 2009,

pp45). My paper will also focus on the comparison of different word representation techniques and investigate into the prediction power of different approaches.

3 Data

3.1 Data Source

To investigate and predict the effect of financial news on stock market trend, news text data and stock market index are needed. News text data used in this project is scraped from Reddit WorldNews Channel and Dow Jones Industrial Average (DJIA) data, which is used to represent stock index, is downloaded from Yahoo finance.

Reddit is an American social news website where users can submit posts, upvote or downvote a post and comment on other people's posts. The sites' content is divided into different communities known as different 'subreddits'. WorldNews Channel is a subreddit where people can post news around the world. The data scraped contains the 25 most popular headlines of news upvoted by users for each given day in the time period from 2008 to 2016.

The Dow Jones Industrial Average (DJIA) is "a stock market index that indicates the value of 30 large, publicly owned companies based in the United States"[1]. The value of DJIA is "the sum of the price of one share of stock for each component company"[2]. The DJIA data from 2008 to 2016 is downloaded from Yahoo finance and contains open, high, low, close, adjusted close price and volume data for each trading day. The adjusted close price is adjusted for both dividends and splits and will be used as the measurement for the stock price movement.

3.2 Data Preprocessing

To prepare the data for my prediction analysis, the news text data and DJIA data need to be combined. The top 25 news headlines are aggregated into a single text string for each day and then cleaned using regular expression to remove some garbled characters. To conduct the binary classification, the target needs to be labeled as either upward or downward trend. In this research, the label is "1" if DJIA Adjusted close price rises or stays as the same, "0" if DJIA Adjusted close price decreases.

The whole dataset is split into a training dataset (from 8th, Aug, 2008 to the 31st, Dec, 2015) and a test dataset (from 1st, Jan, 2016 to 1st, July, 2016) for machine learning model training and test.

4 Method

4.1 Converting Text Data into Vectors

As text data cannot be directly fed into machine learning methods, it has to be converted to numeric representations for further processing. In this research, bag- of-words model, n-grams model and TF-IDF model will be used to represent words as vectors.

Bag-of-words model simply counts how many times each word occurs in a document. It is a basic and easy to understand approach for extracting features from text data. However, bag of words model discards word orders and thus ignores the context. Context and semantics would mean a lot in a corpus and the same words arranged differently may bring different meanings. For example, "he is funny" and

1, 2 From WIKIPEDIA Dow Jones Industrial Average at https://en.wikipedia.org/wiki/Dow_Jones_Indust

“is he funny” would have different meaning.

To overcome this limitation of the simple bag-of-words model, n-grams model will also be used to convert the text data. N-grams is a contiguous sequence of words of length n from a given sample of text or speech. When using bag-of-n-grams model, the context of a word is also considered into the model, thus making it more power than the simple bag-of-words model.

TF-IDF is the abbreviation for term frequency–inverse document frequency. Term frequency refers to the count of a term in a document and inverse document frequency measures if a word is common across all the documents. Thus, TF-IDF measures the relative frequency that a word appears in a document compared to its frequency across all documents. Combining these two terms would help filter out some common words that appear in many documents, such as “the”, “a”, “is”.

In this research, a simple bag-of-words model, a bigram model, a trigram model and a TF-IDF model will be used to convert the text data into vectors. Then these vectors would be fed into various machine learning models for prediction analysis. By comparing the prediction accuracy under the same machine learning model, we could compare the effectiveness of these different word representation models.

4.2 Machine Learning Methods

After converting the text data into vectors, machine learning models are applied to conduct the prediction analysis. Logistic regression, support vector machine, random forest and naïve Bayes models are used for training and test.

Logistic regression is one of the most basic supervised learning techniques for

classification and it is also broadly used in natural language processing. It can be used to classify an observation into different classes. In this research, it classifies the label into upward and downward class.

Support vector machine is a popular machine learning algorithm for natural language processing task. As there are typically very high dimensional data and sparse matrix in NLP tasks, different examples would be distributed into distinct areas of the feature space, which is helpful for SVM to get the clear classification hyperplane.

Random forest is an ensemble classifier which combines the estimation of different decision trees. It fits a set of decision trees with different features and different subsamples of dataset, thus effectively avoiding overfitting.

Naïve Bayes algorithm is based on Bayes theorem with a strong assumption. It has been proved to be fast, reliable and accurate in natural language processing and it usually relies on simple representation of document, such as bag of words.

In this research, all these four machine learning methods will be used to train the model and the prediction accuracy will be compared.

5 Results

After combing the four word vectorization methods and four machine learning methods, the achieved accuracies are listed in table 1 below.

	Bag-of-words	Bigrams	Trigrams	TF-IDF
Logistic Regression	0.8360	0.8412	0.8624	0.8174
SVM Linear	0.8254	0.8598	0.8598	0.8307
Random Forest	0.8439	0.8571	0.8519	0.8307
Naïve Bayes	0.8228	0.8254	0.8254	0.8228

Table 1 Prediction accuracy with different word vectorization methods and machine learning methods

From the accuracy in the above table, trigrams model with logistic regression achieved the highest accuracy score of 0.8624. We can also see that overall there's not much difference in terms of accuracy among different machine learning models and word vectorization techniques.

When comparing different word vectorization methods, it is clear that bigrams and trigrams model work better on news text data than the simple bag-of-words model as they take contexts into consideration. When comparing between bigrams and trigrams, neither one outperforms the other in all situations. We could also notice that TF-IDF doesn't achieve as high accuracy as the other methods.

When comparing different machine learning methods, random forest has the best accuracy with bag-of-words model, support vector machine with linear kernel outperforms other methods with bigrams and logistic regression performs best with trigrams. Overall, the accuracy scores achieved by these four models don't vary much.

6 Conclusion

This paper utilized three different word vectorization techniques and four different machine learning models to predict the stock market trends based on news text data. The accuracy achieved from all the models are above 80%, which strongly confirms that there is significant relationship between news and stock market trend and news text data could be utilized to make predictions for stock price movement. After comparison, n-grams model performs better than other word vectorization techniques in terms of prediction accuracy while there's no significant advantage from different machine learning models.

7 Future Work

Future work to improve the prediction accuracy might include gathering more text data from other data sources and applying deep learning models for prediction. In addition, the news-text-based prediction model could be combined into time series model to make prediction for the exact value of stock price.

References

Fama, E., The Behavior of Stock Market Prices, in Graduate School of Business. 1964, University of Chicago.

Malkiel, B.G., A Random Walk Down Wall Street. 1973, New York: W.W. Norton & Company Ltd.

Gidofalvi, G., Using News Articles to Predict Stock Price Movements. 2001, University of California, San Diego: Department of Computer Science and Engineering

Alanyali, Merve and Moat, Helen Susannah and Preis, Tobias, 2013, Quantifying the Relationship Between Financial News and the Stock Market, Scientific Reports Vol 3 3578 EP - https://doi.org/10.1038/srep03578

Engelberg, J. E. and Parsons, C. A. (2011), The Causal Impact of Media in Financial Markets. The Journal of Finance, 66: 67-97. Available at doi:10.1111/j.1540-6261.2010.01626.x

Cutler, David M. and Poterba, James M. and Summers, Lawrence H (1989) What moves stock prices? The Journal of Portfolio Management Vol 15 No.3 P4, available at http://jpm.iijournals.com/content/15/3/4.abstract

J.Y. Campbell, L. Hentschel (1992), No news is good news: an asymmetric model of changing volatility in stock returns, Journal of Financial Economics, 31, pp. 281-318

Gene Birz, John R. Lott (2011), The effect of macroeconomic news on stock returns: New evidence from newspaper coverage, Journal of Banking & Finance, Vol 35, Issue 11, page 2791-2800,

Robert P. Schumaker and Hsinchun Chen (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans. Inf. Syst. 27, 2, Article 12 (March 2009), 19 pages. DOI: https://doi.org/10.1145/1462198.1462204

Uhl M (2014) Reuters sentiment and stock returns. J Behav Finance 15(4):287–298

Yoshihara A, Seki K, Uehara K (2016) Leveraging temporal properties of news events for stock market prediction. Artif Intell Res 5(1):103–110

Ding X, Zhang Y, Liu T, Duan J (2015) Deep learning for event-driven stock prediction. In: International joint conference on artificial intelligence