What is the relationship between news and stock price trends?

Tianxin Zheng

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.

1 Data:

1.1 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.

1.2 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.

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

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.

2 Method

2.1 Converting Text Data into Vectors

As text data can not 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 “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.

2.2 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.

3 Initial Result

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

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 compare between bigrams and trigrams, neither one outperforms the other in all situations. We could also notice that TF-IDF doesn't work as well as the other methods.

When comparing different machine learning methods, random forest has the best accuracy with bag-of-words model while support vector machine outperforms other methods with bigrams and trigrams.