Predicting the Stock Market Movement using Sentiment Analysis of News Headlines

Project Report

	University of South Florida	Data Science Programming ISM 6930
	Student Names	Arundhathi Patil Nidhi Parashar Bibin Jose

Index

1. Acknowledgement

Our team would like to extend our sincere thanks to all individuals who helped us throughout the course of this project. Special thanks to Prof. Balaji Padmanabhan for his guidance and support.

2. Introduction and Motivation

Predicting trends in stock market prices is one among the many complex machine learning problems. Due to the presence of large number of factors which contribute to changes in the supply and demand, the stock market is a highly volatile field. There are many challenges present in predicting stock trends both in the short-term and long-term, which cannot be handled through the traditional forecasting methods. To minimize the risk in investing for all investors, market risk and forecasting errors need to be minimal.

This report is the outcome of experiments that present an approach that combines machine learning models along with the sentiment analysis of public sentiments from online news sources. Utilizing both parts of this approach, we can predict stock trend more accurately. Essentially, we are defining the problem as a classification problem, where the class labels indicate an increase or decrease in the stock market index.

For this purpose, the potential of algorithms such as random forests, support vector machines (SVM) and XGBoosted trees is explored. We use Natural Language Processing (NLP) along with machine learning models, which can have a lasting impact on how information is consumed and produced by the financial markets and investors. Through our project, we arrive at the conclusion that the model which includes sentiment analysis information performs significantly better than a model with financial predictors alone.

The objective is to create a news monitoring and stock movement prediction system that helps the amateur investor who does not have access to sophisticated trading tools.

Existing Literature

There have been many research efforts in the past to use language features in stock market prediction. Xie et al. (2013) demonstrated the use of textual data such as financial news articles about a company and predicting its future stock trend with news sentiment classification [1]. Another example is from Bollen et al. (2010) where Twitter data was analyzed for its sentiments and which were then correlated to the value of the Dow Jones Industrial Average (DJIA) [2].

Previous Research

Early research on stock market prediction relied on two theories; random walk theory and the Efficient Market Hypothesis (EMH) [3].

Efficient Market Hypothesis (EMH)

The underlying principle behind EMH is that all known information about financial instruments such as stocks, is already factored into the prices of those securities. In other words, market prices reflect all available information. This means that although the stock prices are not static and solely driven by new release of information, any such information has already had an impact on that security and cannot be outperformed.

Stock market prices function on the basis of new information, that is news, rather than present and past prices. Since news is unpredictable, stock market prices will follow a random walk pattern and cannot be predicted with more than fifty percent accuracy [4].

Random Walk theory

The random walk theory states that the past movement or price of either a particular stock or the collective stock market is not a good indicator of its future direction. The theory was described by Maurice Kendall in 1953 and essentially it states that stock prices take a random path and regardless of the stock's history and prices there is an equal chance of either the stock going up or down. Many investors still hold this theory to be true but there is evidence both for and against it.

There are some drawbacks to these theories:

- The 2008 financial crash led to a lot of rethinking of existing financial principles and theories. Efficiency fails to explain the speculative economic bubbles and increased volatility. Many critics blame the financial crisis on the belief in the EMH hypothesis which propagated unnecessary reliance on market resilience.
- Behavioral economists concur that the volatile nature of the stock market is the combination of cognitive biases such as overconfidence or information bias which leads to human errors in reasoning. These errors in information processing leads many investors to ignore value stocks and buy growth stocks at higher prices.

With the advent of big data technology and cloud computing, organizations are investing resources to hire experts to analyze large financial datasets and build statistical models. Additionally, deep learning and natural language processing has made the analysis of large amount of publicly available stock information easier and effective.

3. Dataset

News data:

Historical news headlines from Reddit World News Channel (/r/worldnews). They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)

Stock data:

Historical data from https://finance.yahoo.com/quote/%5EDJI/history?p=%5EDJI

(Range: 2008-06-08 to 2016-07-01) Dow Jones Industrial Average (DJIA) data with Date, Open, High, Low, Close, Volume and Adj Close columns. The start trend is added with 0 and 1 value. 1 if the current days open value is greater than previous days adjacent close value.

4. Sentiment Analysis

NLTK

NLP (Natural Language Processing) is a branch of Artificial Intelligence which deals with the interaction of computers with the human languages. This covers various computational methods used by computers to make sense of the human natural language data. It can be as simple as counting word frequencies in a piece of text to comparing different writing styles.

NLTK is an open source python programming toolkit which is extensively being used in processing human language data. It is one of the most successful platforms that provides easy-to-use interfaces with over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries. It is a great toolkit for processing unstructured data like data from online social media sites, Email, Text messages, News and other online sites.

Vader Sentiment Analyzer:

Sentiment Analysis is a process of analyzing a piece of text has a positive, negative or neutral sentiment. Typically there are two types of sentiment analysis approaches: polarity based - in which a piece of text is classified as positive negative or neutral and valence based - where the intensity of the sentiment is taken into account. For example words like 'good' and 'excellent' will be given the same value in polarity based approach, but in valence based approach 'excellent' will be treated as more positive than 'good'.

After preprocessing the data, we calculated sentiment of the news using the Valence Aware Dictionary and Sentiment Reasoner (VADER) which is a type of sentiment analyzer based on lexicons of sentiment-related words. It is especially useful for sentiments expressed in social media such as Twitter, Facebook and News Data. This gives four sentiment metrics for each piece of text that is passed through it - positive, negative, neutral and compound. All these values lie between range of -1 to 1. The compound value is the sum of positive, negative and neutral values and then standardized to range between -1 to 1. For each headline, sentiment intensity is calculated using VADER algorithm, which returns positive, negative, neutral and compound values for the headline. The magnitude represents the intensity.

5. Methodology and Analysis

Experiment

We used two types of data to implement this algorithm: top 25 headlines voted by users in the subreddit group and historic Dow Jones Industrial Average (DJIA) market index for the last 8 years. In our experiments, we first conducted the sentiment analysis using the Vader python package. The resulting compound sentiment scores are used as features to train the classifiers.

The machine learning models we are using in this project can be divided into two categories:

Supervised learning algorithms:

· Logistic regression

· Support Vector Machine

· Random Forests

· Naïve Bayes clustering

· Decision Trees

Unsupervised learning models:

· K-Means Clustering

In the following section, we describe two important models, SVM and XGBoost.

Support Vector Machines

The SVM algorithm can be used for both classification and regression tasks. In SVM, each data point is plotted in an n-dimensional space, with each dimension corresponding to a feature. During training, SVM algorithm separates these labeled data points with a hyperplane. In testing, the class of each new datapoint is determined by plotting and verifying which side the point lies with respect to the hyperplane.

XGBoost

Boosting is the process in which many weak predictors are combined to make a resulting strong classification. It is an ensemble model where new models are sequentially added that predict the residuals of the preceding models which are added to make the final prediction.

Table 5.1 displays the variables in the dataset.

Name	Definition
Open Price	The first price of a stock traded at the beginning of a specified trading day.
Close Price	The last price of a stock in the last transaction on a specified trading day.
Adjusted Close Price	The close price adjusted based on the reflection of dividends and splits.
High	The highest price when a stock traded on a specified trading day.
Low	The lowest price when a stock traded on a specified trading day.
Volume	Total amount of shares or contracts of a stock traded on a specified trading day.
Label	"1" when DJIA Adj Close value rose or stayed as the same; "0" when DJIA Adj Close value decreased.
Start Trend	Comparison of previous day's adjusted closing value and current day's opening value; 1 if the difference is high else 0 if its lower

Split Dataset

The dataset was split using two different approaches. First, was a chronological split using the first six years in the training set and the remaining two years in the test set. Next, we took the first 80% of dataset as the training set, and the rest 20% as the testing set.

Method

The sentiment compound value is used as the input to the ML model for training, whereas the target is a Boolean vector representing rise or fall of stock price.

Table 5.2 describes the variables after the sentiment analysis using Vader package

Column	Metric
Top News # Compound	Compound score as sum of all lexicon ratings standardized between 1- and 1
Top News # Positive	Positive sentiment
Top News # Negative	Negative sentiment
Top News # Neutral	Neutral sentiment

Next, we tried to correlate the sentiment and stock trend to find whether it had any impact on the market. For example, any negative news about Apple can have an impact on the stock prices for Samsung. The challenge is although the sentiments can help reveal whether the news is positive or negative, it is more complicated to capture its effect on the stock price. Therefore, such analysis was reduced to the problem of predicting the direction of the stock price, whether it falls or rises.

6. Results

The accuracy comparisons between the training on the combined sentiment and stock data are displayed in the table.

Case I: Taking data from 2008-08-08 to 2014-12-31 as train and 2015-01-01 to 2016-07-01 as test.

Algorithm	Accuracy with sentiment (%)	Accuracy with start trend and sentiment
Logistic Regression	51.2	59.86
KNN	50.17	53.00
Decision Tree	48.54	52.05
Gaussian Naive Bayes	50.08	61.49
SVM	53.08	62.17
Random Forest	51.45	56.68
XGBoost	51.02	57.20

Case II: Taking 80% of data as train and 20% as test.

Algorithm	Accuracy with sentiment (%)	Accuracy with start trend and sentiment
Logistic Regression	50.0	63.31
KNN	46.98	57.03
Decision Tree	50.75	54.52
Gaussian Naive Bayes	51.25	61.80
SVM	50.75	66.33
Random Forest	46.98	57.78
XGBoost	48.49	62.31

7. Conclusions and Future Work

In this project, we observed that adding textual information from news and performing sentiment analysis combined with the stock price data can greatly enhance the model accuracy. For future, we see an extensive scope of enhancing this technique using deep learning framework. In addition, we can improve the accuracy by predicting how much influence the sentiment data has on the stock price movement. For this, a regression model can be developed. In addition to sentiment analysis of social media, other indicators such as policy changes, can be used as qualitative indicators to predict price trend. Some other types of data can also be considered like the Financial news data and social media data like Tweets related to financial news or stocks to see if there is an increase in the prediction accuracy of models.

8. References

[1] Boyi Xie, Rebecca J. Passonneau, Germn Creamer, and Leon Wu. 2013. Semantic frames to predict stock price movement. In Proceedings of the 2013 Annual Meeting of the Association for Computational Linguistics.

[2] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. CoRR, abs/1010.3003.

[3] Fama, E. F. The Behavior of Stock-Market Prices. Journal of Business, 1965; 38(1), 34105.

Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved 03/25/2019 from https://www.kaggle.com/aaron7sun/stocknews

https://www.kaggle.com/lseiyjg/use-news-to-predict-stock-markets

http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html