

CSE Honors - I Report

Stock Prediction using Sentiment Analysis from News Headlines

Freya Mehta (20171184)

18th December, 2019



Contents

1. Project Description
2. Introduction
3. Sentiment analysis of news feed
4. Stock Prediction
5. Paper Summaries
6. Various Methods to accompany
sentiment analysis on stock data
7. Model
8. Results
9. Datasets Explored
10. References

Project Description

Predict an individual company's future stock prices (AAPL) using sentiment analysis along with ENN and ANN.

Sentiment analysis is a method to understand public attitude towards a topic/product (here stocks).

Approach - Doing sentiment analysis (mining) on the news feeds play a major role in stock price prediction. Collect thousands of news headlines. Predict if the headline is bullish or bearish. Polarise each data set into four categories, namely compound, positive, negative and neutral for text sentiment analysis. Use this set of data to predict the positivity of current news and effectively predict the stock price.

Introduction

Stock market price prediction is always an important issue and also, one of the most trending topics in the sector of finance, in which many researchers have played their part in order to obtain more accurate results. As, it has great importance in financial gain, therefore it becomes a subject of interest for great many investors, financial analyst, academic and business side people. Predicting the best time of buying or selling is one of the most difficult tasks. Accurate prediction can help investors to acquire more opportunities of gaining profit in the stock exchange. Hence, precise prediction of the trends of the stock price index can be extremely advantageous for investors. However, the behavior of stock markets is based on a great many factors such as previous open, close, volume, low, high and some political events, trader expectations, general economic conditions.

For the prediction of stock market trend, there are two standard measures, namely, technical analysis and fundamental analysis. Technical analysis is a trading tool employed to evaluate securities and identify trading opportunities by analyzing statistics gathered from trading activity, such as price movement and volume. Whereas, Fundamental analysis is the examination of the underlying forces that affect the well-being of the economy, industry groups, and companies.

Sentiment Analysis of News Feed:

For sentiment analysis with NLTK (Natural Language Tool Kit Library) we use VADER library (Valence Aware Dictionary for sEntiment Reasoning is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.) [7]. This library polarized news into four categories, namely compound, positive, negative and neutral. The Positive polarization means, it has a positive effect on stock markets. On the other hand, if the negative polarization means, this tends to have a negative impact on stock markets and makes the stock market values go down. The sentiment of any text can be classified into three major categories (positive, negative, and neutral), the problem is clearly a type of classification. And, we will be using labelled data (supervised learning).

- `SentimentIntensityAnalyzer().polarity_scores('At least nine people were killed in the first strike when missiles destroyed a moving vehicle in the North Assam tribal region, the officials said.')`
{'neg': 0.305, 'neu': 0.695, 'pos': 0.0, 'compound': -0.8481}
- `sid.polarity_scores('Transporters threaten protest hike in fares')`
{'neg': 0.535, 'neu': 0.465, 'pos': 0.0, 'compound': -0.5574}

Stock Prediction:

A stock market is also known as the equity market. It is a facility where traders can buy and sell securities such as shares, bonds and other financial instrument. In simple words stock market trade is the transfer of money of stock or security from a seller to a buyer or vice versa. This requires two parties to confess on a same price. Equities (stocks or shares) consult is an ownership interest in a particular company. Participants in the stock market range can be from small individual stock investors to larger trader's investors, anywhere in the world. Its movement prediction is one of the most difficult task which is influenced by many external factors of social, political and economical. So, the main objective of this experimental study is to improve the prediction accuracy of stock closing price by using sentiment analysis along with ANN and ENN model. For that, we focus on effective indicators that can be used to predict the output variable. We correlate the variable of sentiment analysis with closing price variables of AAPL stock market shares which we get from the daily stock market index. In this project, I used a genetic algorithm (GA) ENN model which is employed to improve the prediction accuracy.

Paper Summaries (Literature Survey)

- Title: Predicting stock using microblog mood

Authors: Danfeng Y, Guang Z, Xuan Z, Yuan T, Fangchun Y.

Year: 2016 (September)

Publisher: IEEE

Summary: SVM and Probabilistic Neural Network were used to make a prediction, and experiments show that SVM is better to predict stock market movements than Probabilistic Neural Network. Individuals' market behaviors will be affected by their own mood states, and the public mood state is reflected by an individual's mood state. This paper mainly discuss stock prediction methods and whether public moods, which are extracted from sentimental analysis of Microblog feeds, can be used to predict stock market movements. There are 2 main tasks: how people express the public moods, and how to predict market movements. First, C-POMS (Chinese Profile of Mood States) was proposed to analyze sentiment of Microblog feeds. Then Granger causality test confirmed the relation between C-POMS analysis and price series.

Sentiment Analysis: It gives a score and divides the text, for example, a piece of Microblog feed, to 3 levels: positive, neutral and negative. Of positive and negative texts, there are still 3 levels: common, medium, and heavy. Then we sum up the total score for each day as the mood index of the day. Profile of Mood States (POMS) is psychological rating scale used to assess transient, distinct mood states. C-POMS includes 7 moods: angry, panic, nervous, energetic,



fatigue, depression and esteem.

In this paper, Granger Causality Test is applied to find the correlation between public moods and stock price series.

- Title : Financial stock market forecast using data mining Techniques

Authors: K. Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam

Year: 2010

Publisher: Proceedings of the international multiconference of engineers and computer

Summary: This paper used data mining technology to discover the hidden patterns from the historic data that have probable predictive capability in their investment decisions. The prediction of stock market is challenging task of financial time series predictions. There are five Methods namely Typical price(TP), Bollinger bands, Relative strength index (RSI), CMI and MA used to analyzed the stock index. In this paper the author got the profitable signal is 84.24% using Bollinger Bands rather than MA, RSI and CMI.


- Title: A Neural network-based fuzzy time series model to improve forecasting

Authors: Tiffany Hui-Kuang yu and Kun-Huang Huarng

Year: 2010

Publisher: Elsevier

Summary: This paper used neural network because of their capabilities in handling nonlinear relationship and also implement a



new fuzzy time series model to improve forecasting. The fuzzy relationship is used to forecast the Taiwan stock index. In the neural network fuzzy time series model where as insample observations are used for training and out-sample observations are used for forecasting. The drawback of taking all the degree of membership for training and forecasting may affect the performance of the neural network. To avoid this take the difference between observations. These reduce the range of the universe of discourse.

- Title: Stock Market forecasting using Hidden Markov Model: A New Approach

Authors: Md. Rafiul Hassan and Baikunth Nath

Year: 2005


Publisher: IEEE 2005

Summary: This paper used Hidden Markov Models (HMM) approach to forecasting stock price for interrelated markets. HMM was used for pattern recognition and classification problems because of its proven suitability for modeling dynamic system. The author summarized the advantage of the HMM was strong statistical foundation. It"s able to handle new data robustly and computationally efficient to develop and evaluate similar patterns. The author decides to develop hybrid system using AI paradigms with HMM improve the accuracy and efficiency of forecast the stock market.

- Title: A hybrid model based on rough set theory and genetic algorithms for stock price forecasting

Authors: Ching-Hsue cheng, Tai-Liang Chen, Liang-Ying Wei

Year: 2010



Publisher: Elsevier

Summary: This paper proposed a hybrid forecasting model using multi-technical indicators to predict stock price trends. There are four procedures described such as select the essential technical indicators, the popular indicators based on a correlation matrix and use CDPA to minimize the entropy principle approach. Then use RST algorithm to extract linguistic rules and utilize genetic algorithm to refine the extracted rules to get better forecasting accuracy and stock return. The advantage was discovered that produce more reliable and understandable rules and forecasting rules based on objective stock data rather than subjective human judgments.

- Title: A type-2 fuzzy rule-based experts system model for stock price analysis

Authors: M.H. Fazel Zarandi, B. Rezaee, I.B. Turksen and E.Neshat

Year: 2009

Publisher: Expert systems with Applications

Summary: This paper used a type-2 fuzzy rule based expert system is developed for stock price analysis. The purposed type-2 fuzzy model applies the technical and fundamental indexes as the input variables. The model used for stock price prediction of an automotive manufactory in Asia. The output membership values were projected onto the input spaces to generate the next membership values of input variables and tuned by genetic algorithm. The type-1 method was used for inference and to increasing the robustness of the system. This method was used to robustness, flexibility and error minimization. It is used to forecast more profitable trading in stock markets.

Various Methods to accompany sentiment analysis on stock data:

I read upon the following methods:

- Principal Component Analysis
- Genetic algorithm
- Decision Trees
- These manage to filter out near 80% unwanted representative features.
- HMM
- Artificial Neural Network
- Apply SVM to construct the prediction model and select gaussian radial basis function as kernel function.
- Probabilistic Neural Network based on the DDA (Dynamic Decay Adjustment) method on labeled data using Constructive Training of Probabilistic Neural Networks as the underlying algorithm.
- Evolutionary Neural Network

Model

Dataset of Interest:

Google Finance AAPL stock data from 21/12/2009 to 31/12/2014

In my project, I use news headlines to predict the predict trend of Apple Inc. Because Apple is a hardware company, its stock price and its core product have a tight connection. Daily stock price of Apple has been acquired from Google Finance. The acquired data covers the daily OHLC(open-high-low-close) statistics over a period from 21/12/2009 to 31/12/2014. We had originally planned to use the Python Yahoo Finance Api, but it had been deprecated. Instead, Pandas DataReader Module was utilized to obtain Apple's stock price over our desired time span. The stock price gave us a good approximation of Apple's business well-doing over the course of five years.

Financial News Dataset from Reuters

I initially scrapped json from Reuters and cleaned data to get top single headlines with timestamp. This dataset was compiled and first used in [Ding et al. (2014)] (<http://emnlp2014.org/papers/pdf/EMNLP2014148.pdf>), [Ding et al. (2015)] (<https://www.ijcai.org/Proceedings/15/Papers/329.pdf>) and [Ding et al., 2014] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Using structured events to predict stock price movement: An empirical investigation. In Proc. of EMNLP, pages 1415–1425, Doha, Qatar, October 2014. Association for Computational Linguistics. I wrote a python script to arrange the raw data into a cleaner CVS file.

Pre-processing

The dynamics of Stock market price data is not complete understandably as it is closed during weekends and public holidays, when it does not function. We join the real time news scrapped by scrapper with the stock closing price of the same day. We have ignored the gap of missing data, i.e., weekend holidays.

Training

From the above extracted factors that we were extracted using the sentiment analysis are fed to the regression and trained using Artificial neural network (ANN) and Evolutionary Neural Network (ENN) model.

Artificial neural network (ANN) model:

In neural network, the data flows forward to the output continuously without any feedback. I have used a typical four-layer neural network model for predicting the closing price of AAPL stock shares. The input nodes consist of technical variable, while the output layer provides the predicted result based on sentiments variable and true value (actual closing price). Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes. For calculation equation can be written as Eq.

$$I_j = \sum_i w_{ij} O_i + \theta_j,$$

Where, i is the number of neuron nodes, j is the number of hidden layer neurons, w is weights, and O is input neurons, θ is the bias value of hidden nodes. Sigmoid Function is used as activation function to calculate the output of the neuron of our hidden layer or output layer. From available sigmoid function namely, logistic and hyperbolic tangent, Rectifier Linear Unit (ReLU), we use logistic function, as it takes a value between 0 and 1.

$$f(x) = \frac{1}{1 + \exp(-x)}$$

Evolutionary Neural Network (ENN) model:

We apply GAs in the form of ENN to evolve the weights between neurons in different layers in the neural network. Steps for evolving is described below:

Step1- Encoding:

Each gene is presented by weight, connected neurons of different layers. Chromosome consists of the combination of weights and biases as shown in figure -2. Our ENN model architecture is based on four neurons in input layer, 5 neurons in two hidden layers and 1 neuron in the output layer. The first gene of chromosome is w_{15} as shown in figure 2. We have used a real number form to represent the connection weights.

Step 2- Generate the initial population:

Initially population is generated randomly. Each of Initial weights are randomly generated between -2 and 2, and biases are between -1 and 1.

Step 3 - Calculating the fitness values:


As regards the fitness function, we have selected the root mean squared error (RMSE) over a training data set. it's the Eq can be written as:

$$RMSE(C_j) = \sqrt{\frac{1}{N} \sum_{i=0}^n (Y_i - P_i)^2}$$

Where Y_i is the actual value and P_i is the output value of i th training data obtained from the neural network using the weights coded in j th chromosome. Where (C_j) and N is the number of training data.

Step 4 – Selection mechanism:

As, in ENN every time we perform crossover and mutation, a new child is created. In order to make predictions more simpler and easy, we find the best individual from population. We use binary tournament selection scheme. In this every time we select two individuals randomly for the selection of one parent that could generate new offspring by using genetic



operators. After comparing the fitness of these two selected individuals, only best individual could get selected same process is done for the selection of second parent and by that they are able to make offspring or child that could survive for the next generation.

Step 5 - Genetic operators:

We have performed two-point crossover along with one-point mutation if required.

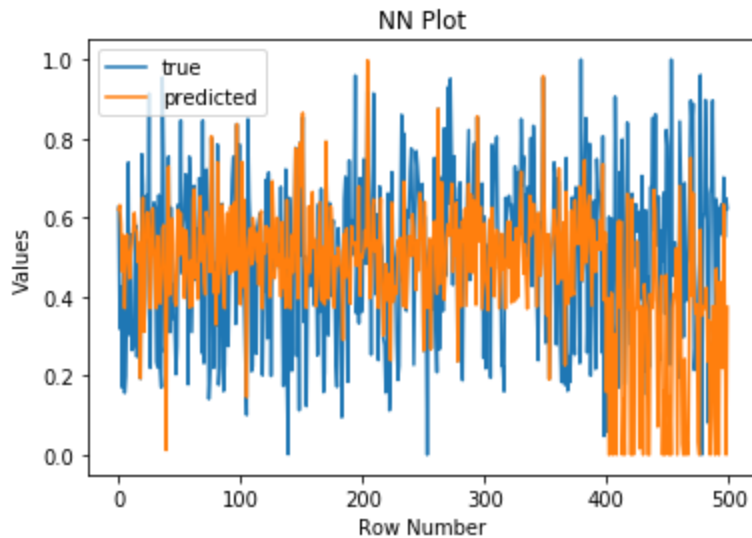
Step 6 –Replacement

Replace the current population by the top fittest that could be used for creating further offspring's.

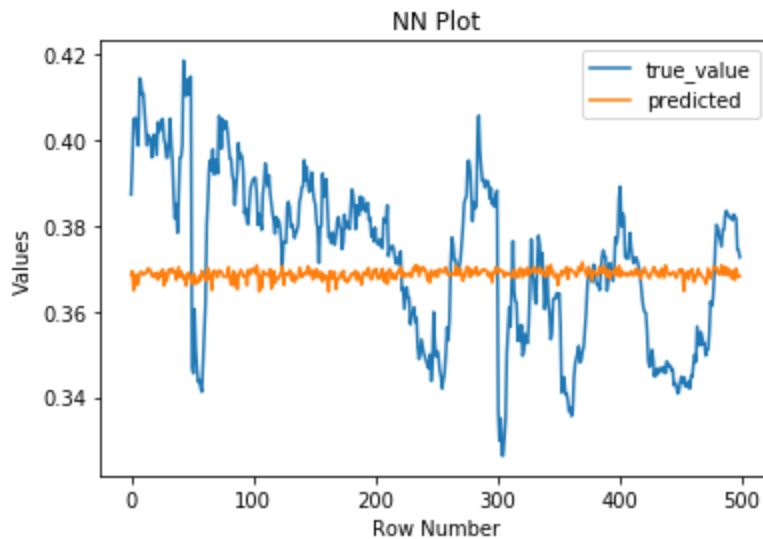
Step 7 - Stopping criteria

If the number of generations equals to the maximum generation number or optimal criteria, then stop; otherwise, go to step 3.

Results




<= ANN Model



<= ENN Model

These results concludes that through an artificial neural network (ANN) model and an evolutionary neural network (ENN) model we can predict the variable value, but an ANN model for prediction of the closing price is far better than ENN model. I used python libraries of pandas, sklearn, numpy



and for accuracy use model R2 score for calculating accuracy of prediction. I have scraper for pre-processing news, data and perform transformation (to improve prediction accuracy through ENN). The results show that the proposed approach of ANN model is able to cope with closing value of AAPL stock shares and it also yields good prediction accuracy in case of closing price prediction as compared to ENN.

Datasets Explored

- Dow Jones Industrial Average (DJIA) - The data was obtained using Yahoo! Finance and includes the open, close, high and low values for a given day.
- BSE Stock Prices:
<https://www.bseindia.com/markets/equity/EQReports/StockPrcHistori.aspx?expandable=6&scripcode=512289&flag=sp&Submit=G>
- News Aggregator Dataset - Headlines and categories of 400k news stories from 2014:
<https://www.kaggle.com/uciml/news-aggregator-dataset/data>
- <https://archive.ics.uci.edu/ml/datasets/News+Aggregator>
- RedditNews.csv: two columns The first column is the "date", and the second column is the "news headlines". All news are ranked from top to bottom based on how *hot* they are. Hence, there are 25 lines for each date. <https://www.kaggle.com/aaron7sun/stocknews>
- US Financial News Articles: Financial News articles available in JSON, set of 306,242 articles. Excellent for text analysis and combined with any other related entity dataset, it could give some astounding results.
<https://www.kaggle.com/jeet2016/us-financial-news-articles>

References

[1] Gholamian Gonabadi D, Mohseni Taheri SD, Mohammadi A, Menhaj MB, editors. Investigating the performance of technical indicators in electrical industry in Tehran's Stock Exchange using hybrid methods of SRA, PCA and Neural Networks. Therm Power Plants IEEE 2014;2014:75–82.

[View Article](#) [Google Scholar](#)

[2] Leung MT, Daouk H, Chen A. Forecasting stock indices: a comparison of classification and level estimation models. Int J Forecast. 2000;16(2):173–190.

[View Article](#) [Google Scholar](#)

[3] Mostafa MM. Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait. Expert Syst Appl. 2010;37(9):6302–6309.

[View Article](#) [Google Scholar](#)

[4] <https://www.investopedia.com/terms/t/technicalanalysis.asp>.

[5] http://stockcharts.com/school/doku.php?id=chart_school:overview:fundamental_analysis.

[6] <https://www.dawn.com/archive/>

[7] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014