Stock Prediction Using Twitter Sentiment Analysis - Anshul Mittal, Arpit Goel

Abstract:

In this paper we use twitter data for predicting stock market movements. By applying sentiment analysis and machine learning principles we find the link between “public sentiment” and the “market sentiment”. This work is based on Bollen et al’s paper which predicted the stock market with accuracy of 87%.

Introduction:

Over the years Stock market prediction has become one of the active research areas using machine learning. As per the Efficient Market Hypothesis the stock prices are driven by the people’s emotions over a news and follow a random pattern. Even though this hypothesis is accepted widely by the research community governing the market in general.

In this paper they have tested a hypothesis which was based on behavioral economics. The emotion of people impacts one’s decision-making process. This implies there is a direct correlation between the “public sentiment” and “market sentiment”. The sentiment analysis was performed on the twitter data to find the people emotion and its polarity then it was categorized as 4 different classes Calm, Happy, Alert and Kind. The results of analysis and the previous stock values of DJIA were used to predict the upcoming stock movements. This work used a strategy based on Bollen et al’s famous work on stock prediction.

Abbreviations

SOFNN - Self Organizing Fuzzy Neural Networks

DJIA - Dow Jones Industrial Average

EMH - Efficient Market Hypothesis

GPOMS - Google Profile of Mood States

POMS - Profile of Mood States

SVM - Support Vector Machine

EBF - Ellipsoidal basis function

Algorithm:

The author had implemented the technique inspired from the algorithms used by Bollen et al. The raw data were preprocessed and then fed to sentiment analysis to find the polarity. The polarity and the processed values were then fed into SOFNN a model learning framework. which is used to predict the future stock movements.



Data set:

The major data sets used by the author were Dow Jones Industrial Average (DJIA) values from June

2009 to December 2009. The data were acquired using Yahoo! Finance. And the Twitter data from June 2009 to December 2009 containing around 500 million tweets. The data comprised of timestamp, username, and tweet text.

Sentiment analysis:

One of the important parts of the process was Sentiment analysis as its result were used to train the predictive model. The methodology used in this paper were:

1. Word List Generation

The author used Profile of Mood States to develop their own word list. POMS is a psychometric questionnaire which is used to rate one’s current mood by making them to answer 65 unique questions on a 1 to 5 scale. Similar to Bollen et al’s methods the author used SentiWordNet and a standard Thesaurus to extend the commonly occurring synonyms of the base words.

1. Tweet Filtering

The author filtered tweets which contain the words” feel”,” makes me”, ”I’m” or ”I am” in them.

1. Daily Score Computation

The author used use a simple word counting algorithm to find the

score for every POMS word for a given day-

score of a word = #of times the word matches tweets in a day / #of total matches of all words

The Stanford core NLP software was used for word tagging and then using a word’s position in the sentence to find its importance. However, It was observed that this process, besides being extremely slow was not too beneficial.

1. Score Mapping

The score of each word to the six standard POMS states using the mapping techniques specified in the POMS questionnaire. Then mapped the POMS states to our four mood states using static correlation rules (for example, happy is taken as sum of vigor and negation of depression).

Then the results were validated on specific days such as Thanksgiving Day and Michael Jackson’s death.

As shown in Figure 2, the moods show a sharp rise of various mood states on Thanksgiving whereas one day after MJ’s death, there is a sharp decline in happiness.



1. Various moods after Michael Jackson’s death on 25 June 2009



1. Various moods on Thanksgiving day on 26 November 2009

**Figure 2: Cross validation of our sentiment analysis by analyzing moods on some important events**

**Granger Causality**

The p-value is computed using the Granger Causality analysis. The analysis finds how much predictive information one signal has about another over a given lag period. The p-value measures the statistical significance of our result. It is identified that calmness and happiness are most helpful in predicting the DJIA values as per the Granger causality.

**MODEL LEARNINGAND PREDICTION:**

As the relation between the stocks and the moods is nonlinear, the Granger causality cannot be used further. The author used 4 different algorithms Linear Regression, Logistic Regression, SVMs, Self-Organizing Fuzzy Neural Networks. The LIBSVM library is used for SVM and for other 3 algorithm the author implemented them in MATLAB.

The Self Organizing Fuzzy Neural Net-work (SOFNN) is a five layer fuzzy neural network which uses ellipsoidal basis function (EBF) neurons consisting of a center vector and a width vector.

The author used online algorithms for creating SOFNNs as introduced in which neurons are added or pruned from the existing net-work as new samples arrive.

Neural networks have been a very effective learning algorithm for de-coding nonlinear time series data, and financial markets often follow nonlinear trends.

Bollen et al showed the 87% correlation using SOFNNs, and the authors results also concluded that SOFFNs do the best among all other algorithms, giving nearly 75.56% accuracy.

To measure the accuracy the author have developed a technique called K- fold sequential cross validation. The Model was trained on all days up to a specific day and tested for the next k days.