Paper reviewed: The Predictive Influence of Twitter on Stock Market Fluctuation **(94/100)**

By: Tianye Song, ts7fx

Date: Nov. 1st, 2016

1. Identify the problem **(20/20)**
    a. What is the score based on: if this paper has clearly explained:
        i. What is the problem at hand?
        ii. What will solving this problem help a certain audience?
    b. Highlights
        i. This paper has successfully identified the problem at hand. It aims to "investigate the predictive abilities of Twitter on the fluctuations of the stock market". In other words, it tries to answer the question of whether twitter data can be used to predict future fluctuations of the stock prices.
        ii. This paper has illustrated that its problem at hand is worth investigating. It stated that:
            1. Everybody wants to profit from stock trading → the problem is of interest to many
            2. NYSE is a focal point of the stock trading + Machine learning and social media are the new favorites of the field of stock price prediction
            3. As a result, to build predictive machine learning models using NYSE trading data to try to predict stock prices can be interesting to stock traders, individual stock buyers, and companies.
        iii. This paper has stated the envisioned impact. "The envisioned impact of transitioning to this alternative situation is that **investors would make more profit and reduce financial losses as a result of poor market predictions and decisions**".
        iv. This paper has discussed what data should be used. It will gather stock data from readily available data source OL, and twitter data using twitter API.
    c. Deficiencies
        i. It seems to the writer of this report that the paper being reviewed has met all requirement in this first section of the evaluation. Thus no deficiencies are noted here.
2. Define objectives and metrics **(15/20)**
    a. What is the score based on: if this paper has clearly stated its objectives and performing metrics
    b. Highlights

i. This paper stated that both logistic and linear regression models will be created to try to predict the changes in stock prices. Logistic model will be used for predicting up/down whereas linear regression model will be used for predicting to what extent does the stock price go up/down.

ii. This paper hopes that the analysis will show how market trends may relate to real time social perspective and how social actions may affect stock values.

c. Deficiencies

i. This paper should have included performance metrics description.

3. Understanding the state-of-art **(19/20)**

a. What is the score based on: if this paper demonstrate that it has clear understanding of past work in the topic domain. Additionally, how to approach the problem at hand from a novel angle.

b. Highlights

i. This paper has discussed multiple techniques applied previously on the problem, including decision trees, neural networks, clustering, etc.

ii. On the techniques being reviewed, this paper has discussed what such techniques have accomplished, as well as the deficiencies on such techniques.

iii. This paper has also criticized different machine learning techniques and analyzed each of these technique's strength and weakness when applied to the problem at hand.

c. Deficiencies

i. Clustering is an unsupervised learning method, which fundamentally attempts to group data based on certain attributes from certain feature spaces. This paper's analysis on clustering did not explain well the leap from data grouping to prediction. How can clusters of data help in making predictions?

4. Define hypotheses and approach **(20/20)**

a. What is the score based on: the quality of hypotheses and experimental design.

b. Highlights

i. This paper's hypothesis is: it is believed that a model can be built to predict these stock prices based off text within tweets and those tweets which are written while trading is open in the market will be a stronger predictor than those written while the stock market is closed.

ii. This paper's hypothesis is predictive in that a prediction of stock prices can be made. It is testable in a finite time, since by simply plugging in historical data can test the validity of the hypothesis. It is falsifiable because it can be

the case that tweets have no effect on stock prices. It is novel because this research looks at company handles instead of mood vocabularies. It does lead directly to data, method and experiments.

   iii. This paper has discussed its data preprocessing approach in a fair amount of details. It has also discussed some of the data source's shortcomings.

   iv. This paper has stated both logistic and linear regression models will be used as methods. Both novelty and assumptions of these methods were discussed

   v. This paper has decided to use performance metrics of the regression models, in addition to a ROC curve as the evaluation methods.

   vi. Both regression models evaluated in this paper measured whether predictors, in this case company handle words, are significant predictors to stock prices.

   c. deficiencies

5. Execute approach and report results **(20/20)**
   a. What is the score based on: whether this paper shows adequate amount of understanding as well as critical analysis of its results.
   b. Highlights
      i. The paper has discussed its result in a critical way. It has rejected its hypothesis, claiming that "hypothesis that the tweets made during the open market period would be better predictors of stock market change has been disproved".

      ii. It has discussed shortcomings in its data sources, as well as its regression models.

      iii. It has discussed how its work, despite its hypothesis being rejected, can be used to pave the road for future researches.

      iv. It has described how its results can be useful for certain audience.
   c. deficiencies