e-PGPathshala
Subject : Computer Science
Paper: Data Analytics
Module No 22: CS/DA/22 - Data Analytics - Mining Streams – Real Time Application
Quadrant 1 – e-text

1.1 Introduction

The aim of this chapter is to gives some details about the real time application of stream mining. The case study about the stock prediction from tweets is summarised here.

1.2 Learning Outcomes

- Case study about real time application of stream mining
- Understand the stock prediction from tweets

1.3 Sentiment analysis

Sentiment analysis (also known as opinion mining) is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Otherwise it refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. In essence, it is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention.

Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organisations across the world.

1.4 Stock Prediction using social media data

The case study reported here is an adaptation from the paper titled “Stock Prediction Using Twitter Sentiment Analysis “ by Anshul Mittal and Arpit Goel**.** This case study, deals with hypothesis based on the premise of behavioural economics, that the emotions and moods of individuals affect their decision making process, thus, leading to a direct correlation between ”public sentiment” and ”market sentiment”. We perform sentiment analysis on publicly available Twitter data to find the public mood and the degree of membership into 4 classes - Calm, Happy, Alert and Kind (somewhat like fuzzy membership). These moods are used with Dow Jones Industrial Average (DJIA) values to predict future stock movements and then use the predicted values for portfolio management strategy.

This case study uses sentiment analysis and machine learning principles to find the correlation between ”public sentiment” and ”market sentiment” from tweets, then predict public mood and use the predicted mood and previous days’ stock market index values to predict the stock market movements.

1.4.1 Data set used

Two datasets are used for this case study:

- Dow Jones Industrial Average (DJIA) values from June 2009 to December 2009
  - The data was obtained using Yahoo! Finance and includes the open, close, high and low values for a given day
- Publicly available Twitter data containing more than 476 million tweets corresponding to more than 17 million users from June 2009 to December 2009.
  - The data includes the timestamp, username and tweet text for every tweet during that period

The entire process and the steps involved are as shown in figure 1. The various steps are explained below:

1.4.2 Data Preprocessing

The data obtained from the above mentioned sources had to be pre-processed to make it suitable for reliable analysis. The pre-processed the DJIA data in the following manner:

1. The Twitter data was available for all days lying in the giving period, the DJIA values obtained using Yahoo. Finance was absent for weekends and other holidays when the market is closed. In order to complete this data, the approximation of the missing values using a concave function. If the DJIA value on a given day is x and the next available data point is y with n days missing in between, we approximate the missing data by estimating the first day after x to be (y+x)/2 and then following the same method recursively till all gaps are filled. This approximation is justified as the stock data usually follows a concave function, unless of course at anomaly points of sudden rise and fall.

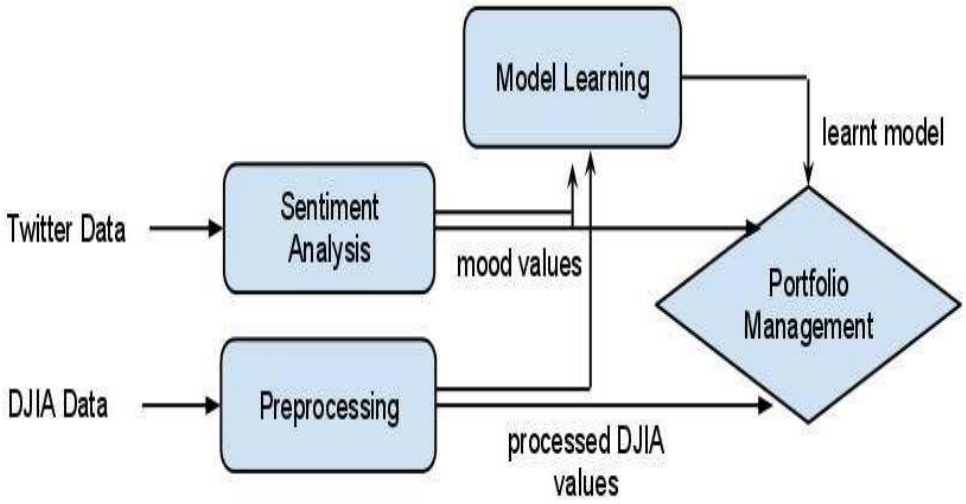

Figure 1. Model for Stock market prediction

2. The general movement of stock markets, it is associated with a few sudden jumps/falls and a brief period of small fluctuations around the new value, such jumps/falls are due to some major aberrations and cannot be predicted. Therefore, the values can be adjusted by shifting up/down for steep falls/jumps, respectively; making sure that won't disturb the daily directional trend (up/down movement of stock prices).

3. Even after shifting the values in step 2, the values contained significant periods of volatile activity which are very difficult to predict. Pruning has done to the dataset by removing these periods for final training and testing.

Finally, in order to ensure that values were small and comparable, we computed the z-score of each point in the data series ((x−μ)/) an used that in our analysis (The original values were of the order of 104, so MATLAB was giving a precision error when computing functions like $exp(-x.^2)$)

1.4.3 Sentiment analysis

Sentiment analysis was an important part of this work since the output of this module was used for learning the predictive model. While there has been a lot of research going on in classifying a piece of text as either positive or negative, there has been little work on multi-class classification. In this case study, there are four mood classes, namely, Calm,Happy, Alert, and Kind. Methodology adopted in finding the public sentiment are

Word List Generation

A word list based on the well known Profile of Mood States (POMS) questionnaire was delveloped. POMS is an established psychometric questionnaire which asks a person to rate his/her current mood by answering 65 different questions on a scale of 1 to 5 (For example,rate on a scale of 1 to 5 how tensed you feel today?).These 65 words are then mapped on to 6 standard POMS moods- Tension, Depression, Anger, Vigour,Fatigue and Confusion. In order to do automate this analysis for tweets, the word list needs to be appropriately extended. By considering all commonly occuring synonyms of the base 65 words using SentiWordNet and a standard Thesaurus.

Tweet Filtering

As mentioned earlier, the tweet data is enormous and will take several hours to be processed if used as it is (which makes the task of daily predictions difficult).Therefore, case study has filtered and considered only those tweets which are more likely to express a feeling, i.e. we consider only those tweets which contain the words ”feel”,”makes me”, ”I’m” or ”I am” in them.

Daily Score Computation

A simple word counting algorithm is used to find the score for every POMS word for a given day-

$$\text{score of a word} = \frac{\#\text{of times the word matches tweets in a day}}{\#\text{of total matches of all words}}$$

The denominator accounts for the fact that the number of tweets could vary from one day to another.

Score Mapping

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

Granger Causality

In order to ascertain whether the mood values returned by the algorithm can be used to predict the future stock movements, they computed the p-values using Granger Causality analysis. Granger Causality 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 i.e. how likely we could obtain the causality value by random chance; there fore, lower the p-value, higher the predictive ability.

1.5 Portfolio management

Having predicted the DJIA closing values one day in advance, use these predicted values to make intelligent sell/buy decisions. Following are the steps/features of this strategy:

- Pre-computation - maintain a running average and standard deviation of actual adjusted stock values of previous k days
- Buy Decision - If the predicted stock value for the next day is n standard deviations less than the mean, we buy the stock else we wait.
- Sell Decision - If the predicted stock value is m standard deviations more than the actual adjusted value at buy time, we sell the stock else we hold.

The causative relation between public mood as measured from a large scale collection of tweets from twitter.com and the DJIA values have been investigated as part of this case study.

	**Other Case Studies**

A)
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852
Predicting

Predicting Elections with Twitter:What 140 Characters Reveal about Political Sentiment

This study uses the context of the German federal election to investigate whether Twitter is used as a forum for political deliberation and whether online messages on Twitter validly mirror offline political sentiment.

B)

http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2826/3237/

Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena

In this article, they perform a sentiment analysis of messages published on Twitter in the second half of 2008. They measure the sentiment of each tweet using an extended version of the Profile of Mood States (POMS) and compare our results to a timeline of notable events that took place in that time period.

Summary

- The analysis doesn't take into account many factors.
  - Firstly, our dataset doesn't really map the real public sentiment, it only considers the twitter using, english speaking people.
- It's possible to obtain a higher correlation if the actual mood is studied.
- It may be hypothesized that people's mood indeed affect their investment decisions, hence the correlation.
- But in that case, there's no direct correlation between the people who invest in stocks and who use twitter more frequently, though there certainly is an indirect correlation - investment decisions of people may be affected by the moods of people around them, ie. the general public sentiment.