

**DMPA LAB PROJECT**

**Stock Analysis**



**Submitted by:~**

**Name:** Nishant Sahoo, Harshini N **Instructor:** Ass. Prof. Chethan Sharma **Reg. No.:** 150953244, 150953250 **Course:** Data Mining and Predictive Analysis Lab Analysis Lab (ICT 3262) **Lab Section**: CCE A2 **Date**: 9th April, 2018

**PROBLEM STATEMENT**

Earlier studies on stock market prediction are based on the historical stock prices. Stock market prices are largely fluctuating. The efficient market hypothesis (EMH) states that financial market movements depend on news, current events and product releases and all these factors will have a significant impact on a company’s stock value. Investing in stocks that might fall is a huge risk.

**INTRODUCTION**

Predicting stock market movements is a well-known problem of interest. With the advent of social media, the information about public feelings has become abundant. Social media is transforming like a perfect platform to share public emotions about any topic and has a significant impact on overall public opinion. Especially Twitter, a social media platform, has received a lot of attention from researchers in the recent times.

Twitter is a micro-blogging application that allows users to follow and comment other users thoughts or share their opinions in real time. More than million users post over 140 million tweets every day. This situation makes Twitter like a corpus with valuable data for researchers. Each tweet is of 140 characters long and speaks public opinion on a topic concisely. The information exploited from tweets are very useful for making predictions.

Stock market prediction on the basis of public sentiments expressed on twitter has been an intriguing field of research. This project aims to observe how well the changes in stock prices of a company, the rises and falls, are correlated with the public opinions being expressed in tweets about that company. Understanding author’s opinion from a piece of text is the objective of sentiment analysis. There are many studies involving twitter as a major source for public-opinion analysis. Google flu trends are being widely studied along with twitter for early prediction of disease outbreaks. This is one of the studies that showcases twitter as a valuable source and a powerful tool for conducting studies and making predictions.

**OBJECTIVE**

To build an accurate model that can predict whether the stock value of a company will increase or decrease depending on the views and public opinions on that company and its products on Twitter.

**METHODOLOGY**

**Data Collection**

A total of 57379 tweets over a period of March 31st, 2016 to June 30th, 2016 on Amazon are extracted. Not only the opinion of public about the company’s stock but also the opinions about products and services offered by the company would have a significant impact and are worth studying. Stock opening and closing prices of Microsoft from March 31st, 2015 to June 30th, 2016 are obtained from Yahoo Finance.

**Data Pre-Processing**

Stock prices data collected is not complete understandably because of weekends and public holidays when the stock market does not function.

Tweets consists of many acronyms, emoticons and unnecessary data like pictures and URL’s. So tweets are preprocessed to represent correct emotions of public. For preprocessing of tweets we employed three stages of filtering: Tokenization, Stopwords removal and regex matching for removing special characters.

1) **Tokenization:** Tweets are split into individual words based on the space and irrelevant symbols like emoticons are removed. We form a list of individual words for each tweet.

2) **Stopword Removal:** Words that do not express any emotion are called Stopwords. After splitting a tweet, words like ‘a’, ‘is’, ‘the’, ‘with’ etc. are removed from the list of words.

3) **Regex Matching for special character Removal:** Regex matching in Python is performed to match URLs and are replaced by the term URL. Often tweets consists of hashtags(#) and @ addressing other users. They are also replaced suitably.

After these stages the tweets are ready for sentiment analysis.

**Sentiment Analysis**

Sentiment analysis task is very much field specific. There is lot of research on sentiment analysis of movie reviews and news articles and many sentiment analyzers are available as an open source. The main problem with these analyzers is that they are trained with a different corpus. For instance, Movie corpus and stock corpus are not equivalent. So, we developed our own sentiment analyzer. Tweets are assigned sentiment values using a python library, Textblob for the same.

**Model Training**

Decision tree classifier is the model used to train the data where a 51 days tweets’ sentiment value is mapped to change in stock for that particular day.

**Correlation Analysis of Price and Sentiment**

The stock price data of Amazon are labeled suitably for training using a simple program. If the previous day stock price is more than the current day stock price, the current day is marked with a numeric value of 0, else marked with a numeric value of 1. Now, this correlation analysis turns out to be a classification problem. The total positive, negative and neutral emotions of all tweets are calculated successively which are used as features for the classifier model and the output is the labeled next day value of stock 0 or 1. The accuracy of the classifier is discussed in the results section.

**RESULTS**

This section gives an overview of accuracy rates of the trained classifier.

1. **Sentiment Analyzer Results**

The above section discussed the method followed to train the classifier used for sentiment analysis of tweets. The sentiment analyzer is not 100% accurate. But it is pretty good in the sense that it gives a higher value for positive tweets and a lower value for negative ones.

1. **Stock Price and Sentiment Correlation Results**

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual Class vs Predicted Class** | **Yes (1)** | **No (0)** | **Total** |
| **Yes (1)** | **TP = 25** | **FN = 18** | **P = 43** |
| **No (0)** | **FP = 4** | **TN = 4** | **N = 8** |
| **Total** | **P’ = 29** | **T’ = 22** | **P+N = 51** |

Accuracy of the model = (TP + TN) / (P + N)

= (25 + 4) / 51

= 0.5686274

The model predicted the correct stock change in nearly 57% of the cases. For example, the sentiment value of tweets for Thursday is 0.1030688 which is pretty good. Using this value, the model predicted that for Thursday the stock values are supposed to go up which is true in reality (confirmed from Yahoo Finance). Similarly, the sentiment value for Friday is 0.051508. The model predicted that for Friday the stock value should go down and again this was true (as confirmed from Yahoo Finance).

The break point for the Decision tree classifier is 0.05614616. Any sentiment value above this will classify it to 1 and below this to 0.

These results give a significant edge to the investors and they show good correlation between stock market movements and the sentiments of public expressed in twitter. This trend shows that with increasing dataset the models are performing well. We would like to incorporate more data in our future work.

**CONCLUSION**

The KDD process in Data Mining is carried out successfully. Textblob library of Python is used to compute the sentiment values of the tweets. We used Decision tree to predict the whether the stock value will go up or down based on the sentiment value of the tweets related to a company. Predictive accuracy is also computed. The model can be improved by using a polynomial regression model and also by training the model with more number of datasets. With increasing size of training datasets, the models tend to perform better.

In this project, we have shown that a strong correlation exists between rise/fall in stock prices of a company to the public opinions or emotions about that company expressed on twitter through tweets. The main contribution of our work is the development of a sentiment analyzer that can judge the type of sentiment present in the tweet. At the beginning, we claimed that positive emotions or sentiment of public in twitter about a company (more sentiment value) would reflect in its stock price. Our speculation is well supported by the results achieved and seems to have a promising future in research.

But we can also conclude that tweets are not the only factor that affect stocks. There are many other factors which also influence the stocks. So, using this model will certainly help a person investing in stocks to be more decisive but one cannot rely totally on this.