Sentimental Analysis of Twitter Feeds for Stock Market Predictions

[1]Priyanka Malpekar,[2]Aishwarya Dudgol,[3]Aparna Sawant,[4]Amit Panpatil,[5]Mrs. Varsha Bodade.

[1,2,3,4] Student, [5]Professor, Information Technology Department.

[1]malpekarpriyanka6@gmail.com,
[2]aishwaryadudgal@gmail.com,
[3]sawantaparna75@gmail.com

Abstract-- **Twitter messages, or tweets, can provide an accurate reflection of public sentiment on when taken in aggregation. We propose a novel approach to label social media text using significant stock market events (big losses or gains). Since stock events are easily quantifiable using returns from indices or individual stocks, they provide meaningful and automated labels. We will extract significant stock movements and collect appropriate pre, post and contemporaneous text from social media sources (for example, tweets from twitter). Subsequently, we assign the respective label (positive or negative) for each tweet. We train a model on this collected set and make predictions for labels of future tweets. We aggregate the net sentiment each day (amongst other metrics) and show that it holds significant predictive power for subsequent stock market movement. Additionally, we apply extracted twitter sentiment to accomplish two tasks. We first look for a correlation between twitter sentiment and stock prices. Secondly, we determine which words in tweets correlate to changes in stock prices by doing a post analysis of price change and tweets. We accomplish this by mining tweets using Twitter's search API and subsequently processing them for analysis.**

Keywords: **Analysis, tweets, twitter, sentiments, Stock, twitter API.**

1. INTRODUCTION

Throughout a day, large amount of texts are transmitted online through a variety of social media channels. Within these texts, valuable information about virtually every topic exists. Through Twitter alone, over 400 million tweets are sent per day. Though each tweet may not seem extremely valuable, it has been argued that the aggregations of large amounts of tweets can provide valuable insight about public mood and sentiment on certain topics. Gauging the public's sentiment by retrieving online information on the market can be valuable in creating trading strategies and for making future stock market predictions. This is not the technical analysis of stocks, but rather fundamental analysis through data mining. There are many factors that are involved in prediction or analyzing the movement of stock prices but public sentiment is included in our project.

In this project, we will be investigating the relationship between the sentiment of tweets to a particular stock and change in the price of the stock over a given period of time. Of course, true stock price correlations involve many more factors, but we will just look at the correlation between Twitter sentiment and stock price, merely as an interesting application of sentiment analysis.

2. LITERATURE SURVEY

There are many studies that aim to identify a method to predict or even understand financial market movements properly. After the popularization of social media, such as blogs, microblogs and sites such as Facebook, there is now a new data source. This data source contains huge amounts of data; therefore, with the help of developed computer technologies, the use of sentiment analysis techniques has grown in recent years.

Analyzing the effects of sentiments on stock performance such as Wysocki (1998), Tumarkin and Whitelaw (2001), Dewally (2003), Antweiler and Frank (2004) have determined, with various degrees of success, the role that sentiments in the form of micro blogs or micro stock message board activities have in the prediction of excessive stock returns. These findings have been supported by Tayal and Komaragiri (2009), who found that analyzing micro-blogs is more reliable for predicting the future performance of a company than larger regular blogs. These studies have found that Twitter sentiments have a significant predictive ability for stock behavior. In a related study of the forex market, Vincent and Armstrong (2010) found that Twitter data are more useful for predicting break points than classical methods. However, their analysis did not attempt to evaluate the predictive impact of Twitter-based sentiments on the market, nor did they analyze the market of an emerging economy, which we will be undertaking in our project.

3. EXISTING SYSTEM

There have been a number of prior works using sentiment analysis to predict stock data. Most notably, Bollen et al. used Twitter data to predict the direction of DJIA movement, achieving an accuracy of 87.6% using the self-developed Google Profile of Mood States (GPOMS) and a self-organizing Fuzzy Neural Network [2010]. They found that the "Calm" mood profile yields the best result for stock market prediction. In similar research, Oh and Sheng found that using a combination of manual and an automated "bag-of-words" technique yielded high predictive accuracy for a variety of stock tickers [2011]. Their SMO algorithm was able to achieve a weighted F-measure of 85.3%.

4. PROPOSED SYSTEM

4.1 Twitter Search/Streaming API

For tweet collection, Twitter provides a rather robust API. API is a set of routines, protocols and other tools for building software applications. There are two possible ways to gather Tweets: the Streaming API or the Search API. The Streaming API allows users to obtain real-time access to tweets from an input query. The user first requests a connection to a stream of tweets from the server. Then, the server opens a streaming connection and tweets are streamed in as they occur, to the user.

However, there are a few limitations of the Streaming API. First, language is not specifiable, resulting in a stream that contains Tweets of all languages, including a few non-Latin based alphabets. Additionally, at the free level, the streamed Tweets are only a small fraction of the actual Tweet body. Initial testing with the streaming API resulted in a polluted training set as it proved difficult to obtain a pure dataset of Tweets.

Because of these issues, we decided to go with the Twitter Search API instead. The Search API is a REST API which allows users to request specific queries of recent tweets. REST, or Representational State Transfer, simply uses HTTP methods (GET, POST, PUT, and DELETE) to execute different operations. The Search API allows more fine tuning of queries, including filtering based on language, region, and time. The query is constructed by stringing separate keywords together with an "OR" in between. Though this is also not a fully complete result, it returns a well filtered set of tweets that is useful for our sentiment analyzer. The request returns a list of JSON objects that contain the tweets and their metadata. This includes a variety of information, including username, time, location, retweets, and more. For our purposes, we mainly focus on the time and tweet text. Both of these APIs require the user to have an API key for authentication. Once authenticated, we were able to easily access the API through a python library called Twython - a simple wrapper for the Twitter API.

4.2 Training Set Collection

Using Twitter's search API, we formed two separate datasets (collections of Tweets) for training: "positive" and "negative." Each dataset was formed programmatically and based on positive and negative queries on emotions and keywords:

- Positive sentiment query: ":) OR :-) OR =) OR :D OR <3 OR like OR love"
- Negative sentiment query: ":(OR =(OR hate OR dislike"

The tweets that contained these keywords or emotions were most likely to be of that corresponding sentiment.

4.3. MongoDB Storage

To save our training tweet data, we used MongoDB. MongoDB is a document based database in which data is stored as JSON-like objects. Each tweet is simply stored as a record in a collection in the database. Only relevant information, including tweet text and date is stored. Querying the MongoDB database is simple.

4.4. Tweet Text Processing

The text of each tweet contains many unwanted words that do not contribute to its sentiment. Many tweets include URLS, tags to other users, unwanted acronyms or symbols that have no meaning. To accurately obtain a tweet's sentiment, we first need to filter the noise i.e the unwanted data from its original state.

A. Tokenization

The first step involves splitting the text by spaces, forming a list of individual words per text. This is also called a bag of words.

B. Removing Stopwords

Next, we remove stopwords from the bag of words. Python's Natural Language Toolkit library contains a stopword dictionary. To remove the stopwords from each text, we simply check each word in the bag of words against the dictionary. If a word is a stopword, we filter it out. The list of stopwords includes articles, some prepositions, and other words that add no sentiment value.

C. Twitter Symbols

Many tweets contains extra symbols such as "@" or ‘‘#" as well as URLs. The word immediately following an "@" symbol is always a username, which we filter out entirely, as they add no value to the text. Words following `‘#" are kept, for they may contain information about the tweets, especially for categorization. URLs are filtered out entirely, as they add no sentiment meaning to the text.

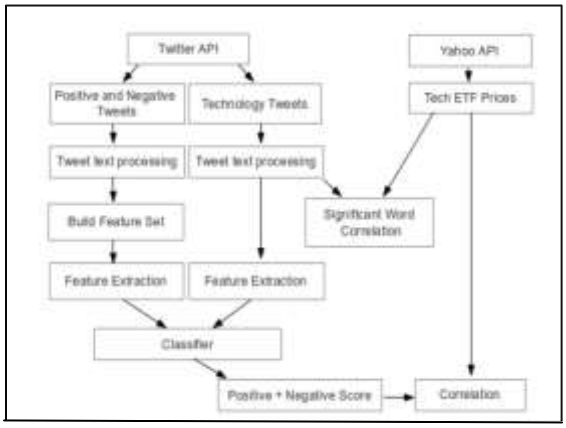


Figure 1: Proposed System

D: Working

1. Tweets will be collected using programming language JSP, servlets and twitter API over a seven week period. These tweets will be collected each week and stored using MYSQL database management system.
2. The tweet text will then be read back into java, pre-processed and read for analysis.
3. For the word feature analysis, a term document matrix will be built and a word frequency and word count will be constructed.
4. For the sentiment analysis, the pre-processed tweets for each day will pass through a sentiment function which will compare the clean text to positive and negative word list and tweets will be scored accordingly. A positive word will receive a '+1' and a negative word will receive a '-1'. The daily total score was then calculated.
5. The data on the activity of the stock will then be downloaded from yahoo finance which includes opening price, closing price, volume and daily changes.
6. Analysis of data will be done for correlation of twitter and yahoo finance data and future stock prices will be predicted.

5. CLASSIFIERS

Accurate classification is still an interesting problem in data mining. Many times, we want to build a classifier with a set of training data and labels. In our case, we want to construct a classifier that is trained on our "positive" and "negative" labelled tweet corpus. From this, the classifier must be able to label future tweets as either "positive" or "negative" based on the tweet's attributes or features. We will be examining three classifiers:
1) Naive Bayes Classifier
2) Support Vector Machine
3) Maximum Entropy

Classifier Evaluation:

We will be calculating precision, accuracy and recall. After examining, we have found that Naïve bayes has the highest accuracy and precision values. Thus we are using Naïve bayes classifier in our project.

In the following example, c will represent the class label, which in our case is either "positive" or "negative," an fi represents a feature in the feature set F.

Naive Bayes:

A Naive Bayes classifier is probabilistic classifier based on Bayes Rule, and the simplest form of a Bayesian network. The classifier is simple to implement and widely used in many applications such as text classification and sentimental analysis. The classifier operates on an underlying "naive" assumption of conditional independence about each feature in its feature set.

The classifier is an application of Bayes Rule:

$$\mathbf{P(c|F) = P(F|c).P(c) \ / \ P(F) \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)}$$

We can treat the denominator, P(F) as a constant, for it does not depend on c. Therefore, we must focus on solving the numerator. To do this, we need to determine the value of P(F|c). Here is where the independence assumption comes in. We assume that, given a class, cj, the features are conditionally independent of each other, therefore:

$$\mathbf{P(f1,f2\ ...fn|cj) = \prod P(fi|cj)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)}$$

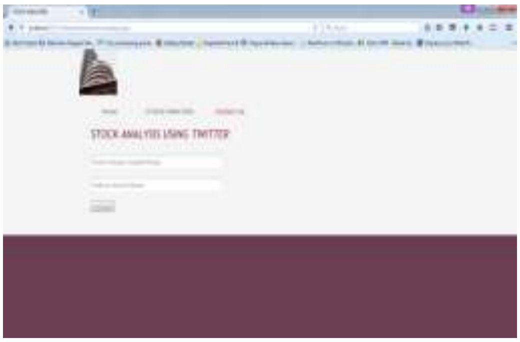
From this, we can classify a tweet with a label c∗ with a maximum posterior decision rule taking the most probable label from all labels C.

$$\mathbf{c* = argmax\ \ P(cj)\prod P(fi|cj)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)}$$

6. ADVANTGES

- The proposed algorithm doesn't generate the candidate itemsets.
- It uses only a single pass over the database.
- The memory consumed is also very less.
- Processing speed is more when compared to rules generated using item set tree and DS theory.
- The proposed system can be used by traders and investors as a reference with their investment strategies. The model will provide investors the direction in, which a stock price will trend. It will help investors as they usually get loss because of unclear investment objective and blind investment.

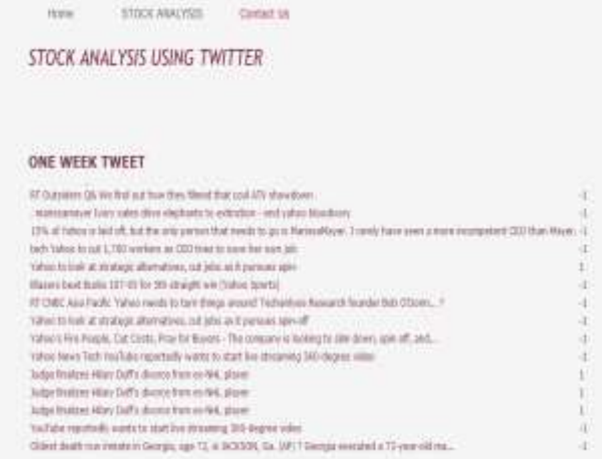
7. SCREENSHOTS



Home STOCK ANALYSIS Contact Us

STOCK ANALYSIS USING TWITTER

Figure 1 : Input page



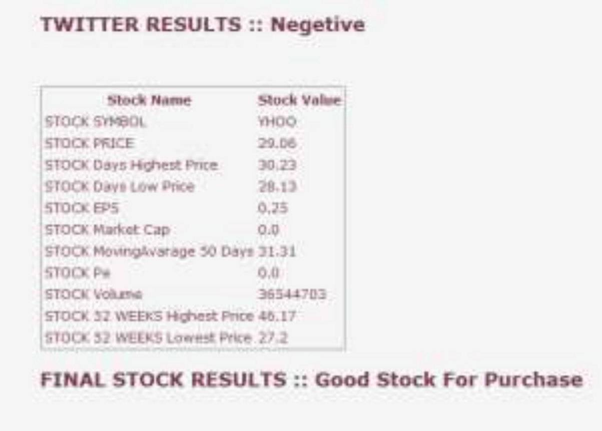
Home STOCK ANALYSIS Contact Us

STOCK ANALYSIS USING TWITTER

ONE WEEK TWEET

RT Outsiders Q& We find out how they filmed that cool ATV showdown -1
marissamayer Ivory sales drive elephants to extinction - and yahoo bloodbath -1
15% of Yahoo is laid off, but the only person that needs to go is MarissaMayer. I rarely have seen a more incompetent CEO than Mayer. -1
tech Yahoo to cut 1,700 workers as CEO tries to save her own job -1
Yahoo to look at strategic alternatives, cut jobs as it pursues spin 1
Blazers beat Bulls 107-95 for 5th straight win (Yahoo Sports) -1
RT CNBC Asia Pacific: Yahoo needs to turn things around! Techandsos Research founder Bob O'Donn... ? -1
Yahoo to look at strategic alternatives, cut jobs as it pursues spin-off -1
Yahoo's Fire People, Cut Costs, Pray for Buyers - The company is looking to slim down, spin off, and... -1
Yahoo News Tech YouTube reportedly wants to start live streaming 360-degree video -1
Judge finalizes Hilary Duff's divorce from ex-NHL player 1
Judge finalizes Hilary Duff's divorce from ex-NHL player 1
Judge finalizes Hilary Duff's divorce from ex-NHL player 1
YouTube reportedly wants to start live streaming 360-degree video -1
Oldest death row inmate in Georgia, age 72, is JACKSON, Ga. (AP) ? Georgia executed a 72-year-old ma... -1

Figure 2: Twitter output page



TWITTER RESULTS :: Negetive

Stock Name	Stock Value
STOCK SYMBOL	YHOO
STOCK PRICE	29.06
STOCK Days Highest Price	30.23
STOCK Days Low Price	28.13
STOCK EPS	0.25
STOCK Market Cap	0.0
STOCK MovingAvarage 50 Days	31.31
STOCK Pe	0.0
STOCK Volume	36544703
STOCK 52 WEEKS Highest Price	46.17
STOCK 52 WEEKS Lowest Price	27.2

FINAL STOCK RESULTS :: Good Stock For Purchase

Figure 3: Final output page

8. CONCLUSION

- We have concluded that even with much simpler sentiment analysis methods, a correlation between Twitter sentiment data and stock market movement can be seen.
- We proposed a novel method for estimating sentiment based on twitter posts and use the sentiment to predict future stock market movement.
- Specifically, we automatically generate training data based on events related with stock markets.
- With such training data, we are able to build a high-precision and efficient classifier to assess tweet sentiment and use such information to build an effective trading strategy.

8. FUTURE SCOPE

- We have investigated the relation between the public moods as measured from large scale collection of tweets from twitter.com.
- Our results show that firstly the public mood can indeed be captured from large scale twitter feeds by means of simple natural language processing techniques.
- There are many areas in which this work could be expanded on in the future. With a longer period of time and more resources, there is much potential in the area.
- Another natural expansion of the project would be to actually make stock predictions and trade money based on twitter sentiment.
- Nevertheless, it would be an interesting area of future study, as the options and opportunities in the area of stock price prediction are endless.

REFERENCES

[1] Sentiwordnet, an enhanced lexical resource for sentiment analysis and opinion mining.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. 2010.

[3] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. 2011.

[4] J. Bollen and H. Mao. Twitter mood as a stock market predictor. IEEE Computer, 2010.

[5] A. Bromberg. Sentiment analysis in python. 2013.

[6] Eric D. Brown. Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. SAIS Proceedings, 2012.

[7] Hal Daum. Notes on cg and lm-bfgs optimization of logistic regression. 2004.

[8] Ray Chen and Marius Lazer. Sentiment analysis of twitter feeds for the prediction of stock market movement. 2012.