Predicting stock price using sentiment analysis combining Twitter, search engine and investor intelligence data

A Thesis

Presented to

The Faculty of the Graduate School

At the University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

By

RUI WU

Dr. Yi Shang, Advisor

DECEMBER 2014

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

Predicting stock price using sentiment analysis combining Twitter, search engine and investor intelligence data

Presented by Rui Wu

A candidate for the degree of

Master of Science

And hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Yi Shang

Dr. Dong Xu

Dr. Jianlin Cheng

**ACKNOWLEDGEMENTS**

It gives me great pleasure to acknowledge the continuous support and guidance of my advisor Dr. Yi Shang throughout the development of my master’s project. His scholarly advice, meticulous scrutiny and patient encouragement have helped me immensely in development of this project. I would also like to thank my lab mate Zhaoyu Li for his assistance and opinions. I would like to thank Dr. Dong Xu and Dr. Jianlin Cheng for being a part of my defense committee and for their invaluable suggestions and comments on this project.

 **Table of Contents**

1. Introduction 6

2. Related Work 8

3. System Architecture 12

4. Methodology 14

4.1 Sentiment Features Generation 14

4.1.1 Filtering and Cleaning 14

4.1.2 Sentiment Analysis Tools 17

4.1.3 Sentiment Features Calculation 18

4.2 Finance Features 19

4.3 Search Engine Query Feature 21

4.4 Investor Intelligence Feature 22

4.5 Correlations between Sentiment features and Finance Features 22

4.6 Machine Learning Algorithms 24

4.6.1 Decision Stump 24

4.6.2 Bootstrap Aggregating 24

4.6.3 Linear Regression 24

4.6.4 Gaussian Regression 25

4.6.5 Radial Basis Function Network 25

4.6.6 Multilayer Perceptron 26

4.7 Present Day/Week Record Generation 26

5. Results and Analysis 27

5.1 Sentiment Feature set 27

5.2 Finance Feature set 28

5.3 Weekly Feature set 29

5.4 Correlations and Analysis 29

5.4.1 Daily Correlations 30

5.4.2 Weekly Correlations 31

5.4.3 Correlations with Time Lag 32

5.5 Machine Learning and Models 34

5.6 Real Time Twitter and Daily Data Collection 38

6. Future Work 40

7. Summary 41

8. References 42

## Abstract

The stock markets in the recent years have become an integral part of the global economy, any fluctuation in this market influences our personal and corporate financial lives. A good prediction model for stock market forecasting is always highly desirable and would of wider interest.

Recent research suggests that very early indicators can be extracted from online social media (blogs, Twitter feeds, etc.) to predict changes in various economic and commercial indicators. In this project, daily sentiment features are generated from a Twitter dataset to build up a high accuracy prediction model for stock price movement. Google Search Queries and Investor Intelligence provide additional features to improve performance on weekly-based models.

Five sentiment features (Mt-Positive, Mt-Negative, Bullishness, Message Volume, Agreement) are extracted from Twitter using sentiment analysis. Tweets that can express opinion upon stocks or indices are filtered out and classified from a Twitter dataset, which holds more than 400 million records from July 31 to December 31 2009. Four finance features (Return, Close, Trade Volume, Volatility) are generated for 2 Market Indices NASDAQ-100, Dow Jones Average Indices and 13 leading technological companies. Second step, correlations on each finance features with all other features are calculated to verify their statistically relationships. Results show high correlations (up to 0.93 for DJIA with Close) with stock prices and twitter sentiment. Twitter Sentiment may have time delay on stock prices movement, so time lag by weeks are also included in this experiments. Furthermore, with confidence from the correlations, several Machine Learning algorithms like Gaussian Process, Neural Network and Decision Stump are applied on the feature set. Results show reliable models are built with strong correlations and low Root Mean Square Error (R: 0.94, RMSE: 0.065). Finally, a real time prediction system is built with an additional component of Twitter Streaming API for collecting real time Twitter data. Overall, the experimental results show that this prediction system is working with satisfiable efficiency and accuracy.

1. Introduction

The stock markets in the recent years have become an integral part of the global economy. The United States Stock market has a market capitalization of US$18.668 trillion (2012) [1]. Movements in the stock market can have a profound economic impact on the economy and everyday people. A collapse in share prices has the potential to cause widespread economic disruption [2]. Most famously, the stock market crash of 1929 was a key factor in causing the great depression of the 1930s. Any fluctuation in this market influences our personal and corporate financial lives. For example, many private pension funds will invest in the stock market. A substantial and prolonged fall in the stock market could lead to a fall in the value of their pension fund, and it could lead to lower pension payouts when they retire. Even if people don’t own shares, it is quite likely people with a private pension will have some connection to the stock market.

Tons of investors are using different type of information resources like news and investor articles to make predictions on stock prices, which could yield significant profit if successful. A good prediction model for stock market forecasting is always highly desirable and would of wider interest.

Numerous attempts have been made to from different aspect. Early study centered on the behaviors of human traders within this socially constructed system. Recently, fundamental and technical analysis techniques have been applied on this prediction problem. One hottest direction is to use online social media like blogs and Twitter. Public sentiments extracted from social media are approved to be an efficient economic and commercial indicator. Another indictor, search engine volume provides valuable reflection of important events at a certain time point. Investors express their opinions by voting on weekly survey. AAII collects their predictions and publishes weekly investor sentiment indicator. In this project, I combine all those three data resources and create a set of features using sentiment analysis. A prediction model is built using real time APIs and models learned from the feature set above.

This paper is arranged as follows. Section 2 provides an overall view of literatures concerning analysis of socioeconomic phenomena using Twitter sentiment, a survey on various data sources predicting finance market, using twitter sentiment to predict stock price movement. Section 3 gives the overall system architecture that can give you’re a pipeline of all the components and data flow. Section 4 describes methodologies including data collecting and processing, feature set generation and correlation analysis, machine learning and new daily/weekly record collection. Section 5 provides experiment results and analysis. Last section is about future work and summary.

2. Related Work

2.1 Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena.

A variegated mosaic of microblogging uses has emerged since the launch of Twitter in 2006: daily chatter, conversation, information sharing, and news commentary, among others. Regardless of their content and intended use, tweets often convey pertinent information about their author’s mood status. As such, tweets can be regarded as temporally authentic microscopic instantiations of public mood state. In this paper [3], authors perform a sentiment analysis of all public tweets broadcasted by Twitter users between August 1 and December 20, 2008. For every day in the timeline, they extract six dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using an extended version of the Profile of Mood States (POMS), a well-established psychometric instrument. They compare their results to fluctuations recorded by stock market and crude oil price indices and major events in media and popular culture, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day. Results show that events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood. The authors speculate that large scale analyses of mood can provide a solid platform to model collective emotive trends in terms of their predictive value with regards to existing social as well as economic indicators.



Figure (1): Spark lines for public mood before, during and after the US presidential election on November 4, 2008.

2.2 Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data

In this paper [4], the authors survey a range of online data sets (Twitter feeds, news headlines, and volumes of Google search queries) and sentiment tracking methods (Twitter Investor Sentiment, Negative News Sentiment and Tweet & Google Search volumes of financial terms), and compare their value for financial prediction of market indices such as the Dow Jones Industrial Average, trading volumes, and market volatility (VIX), as well as gold prices. They also compare the predictive power of traditional investor sentiment survey data, i.e. Investor Intelligence and Daily Sentiment Index, against those of the mentioned set of online sentiment indicators. The results show that traditional surveys of Investor Intelligence are lagging indicators of the financial markets. However, weekly Google Insight Search volumes on financial search queries do have predictive value. An indicator of Twitter Investor Sentiment and the frequency of occurrence of financial terms on Twitter in the previous 1-2 days are also found to be very statistically significant predictors of daily market log return. Survey sentiment indicators are however found not to be statistically significant predictors of financial market values, once we control for all other mood indicators as well as the VIX.

2.3 Analyzing Stock Market Movements Using Twitter Sentiment Analysis

In this paper [5], the authors investigate the complex relationship between tweet sentiment features (like bullishness, volume, agreement etc.) with the financial market instruments (like volatility, trading volume and stock prices). They have analyzed sentiments for more than 4 million tweets between June 2010 to July 2011 for DJIA, NASDAQ-100 and 13 other big cap technological stocks. Our results show high correlation (up to 0.88 for returns) between stock prices and twitter sentiments. Further, using Granger’s Causality Analysis, we have validated that the movement of stock prices and indices are greatly affected in the short term by Twitter discussions. Finally, the authors have implemented Expert Model Mining System (EMMS) to demonstrate that our forecasted returns give a high value of R- square (0.952) with low Maximum Absolute Percentage Error (MaxAPE) of 1.76% for Dow Jones Industrial Average (DJIA).

3. System Architecture



Figure (2): System Architecture

This project can be divided into two part, training the predicting models on feature set generated from 3 different data sources and building a predicting system using the models get from part one. To start with, I collect a whole set of data from Twitter, GIS and IS within the same time period. Five sentiment features , Bullishness, Message Volume, Agreement) are generated from Twitter using sentiment analysis on a daily base. Tweets that can express opinion upon stocks or indices are filtered out and classified from a Twitter dataset. Four finance features (Return, Close, Trade Volume, Volatility) are generated for 2 Market Indices NASDAQ-100, Dow Jones Average Indices and 13 leading technological companies. Stock price raw data is collected from Yahoo Finance. It has daily records of Close, Open, High, Low and Trade Volume. Google Search Queries are manual collect from Google Insights Search on weekly base. Investor Sentiment is collected from The American Association of Individual Investors with investor sentiment of bullish, neutral and bearish, also on weekly base. Second, I calculate correlations on each finance features with all other features. My result shows high correlations (up to 0.814 for Close) with stock prices and twitter sentiment. Furthermore, Machine Learning algorithms like Gaussian Process, Additive Regression and Decision Stump on my feature set. The result shows reliable models are built with strong correlations and low Root Mean Square Error.

Finally, a prediction system is built using models I have above. Real time tweets are collected and stored using Twitter Streaming API. These files go through another filtering and cleaning process and then used to extract new features. GIS and IS data can be get from the same source where training data are collected. Feature generation Daily/weekly sentiment features are always calculated at the end of the day/week and put into the model to get prediction result.

4. Methodology

4.1 Sentiment Features Generation

Sentiment feature set are an important component of training features. Five daily sentiment features of Twitter need to generate from the raw Twitter dataset, which contains more than 400 million records from July 31 2009 to December 31 2009, 154 days. Data size is 58.4 Gigabytes. It is a subset of Twitter dataset released by Stanford SNAP [6], the organization estimate this is about 20-30% of all public tweets published on Twitter during the particular time frame. The data format is not the same as Twitter Streaming API output, which is in JSON format and has more tags like region and languages. In this dataset, Date, User ID and Content have been extracted to consist a full record. The data set from Twitter is complete data without any miss value or non-consistent days, but still, several steps are required to get useful subset to do prediction.

4.1.1 Filtering and Cleaning

The size of twitter raw data is too large for existing machine learning algorithms. And specific tweets that can reflect public opinions on stocks are the only part wanted. Figure 3 shows the workflow of data cleaning and filtering.

Figure (3): Data cleaning and filtering workflow

First, tweets containing each stocks and indices are filtered out by name and stored in separated files. This step can easily reduce file size from gigabytes to megabytes. Each file then goes through a data cleaning step, which conversing the content to lower-case and removing non alphanumeric characters. This cleaning step makes every file ready to go through later filtering. Origin Dataset also contains tweets from all over the world, so it has several languages like Japanese and French. A language Filter library call JLangDetect [8] is also applied as one filter layer. The result then contains only target company related tweets which express opinion and in English. JLangDetect uses parallel corpus and additional languages profiles which enable it to detect 25 different languages [9]. Since I am interested only in those tweets that represent an explicit expression of individual sentiment, or those that reflecting an individual’s present status, I only retain tweets that match the following regular expressions like “feel”, “I’m”, “Im”, “am”, “being”, and “be” [7]. Full list is given in Table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| i am  | im | i’m | am | i feel | i don’t feel |
| feeling | makes me | it seems | opinion | view | it’s |
| it is | say | believe | subject | doubt | consider |
| think | suppose | assume | presume | guess | bet |

Table (1) Opinion expressions list

In order to avoid spam messages and other information-oriented tweets, we also filter out tweets that match the regular expressions ”http:” or ”www.”. The diversity of expression on Twitter is tremendous and few tweets directly express or pertain to users’ mood or sentiment. Tweets conforming to these particular patterns are a minority and consequently this procedure leads to a reduction of the 400 million tweets to a set of 27 thousands (Comparing 9.6 million tweets to 1.1M in preview research [7]) containing mostly expressions of individual mood states. On average, it’s about 207 records each day each company. There are also some special cases. For market indices DJIA and NASQAD, opinion key word filter is removed, since indices are not normally used in life and social occasions. All tweets contains these words should directly related to market activities or opinions expression. The overall goal of this step is to filter out tweets that are most related to stock market movement and as many as it could be on a daily base. Here provides the pseudo code for filtering and cleaning.

*Function FilterByName*

*FOR each tweet IN dataset*

 *IF tweet#content CONTAINS word IN names*

 *THEN WRITE tweet TO file(name);*

*END*

*Function Cleaning&Filtering*

*FOR each file*

 *FOR each tweet IN file*

 *CALL toLowerCase(tweet#content);*

 *CALL removeNonAlphaChars(tweet#content);*

 *IF langDect(tweet#content) equals 'EN'*

 *AND tweet#content CONTAINS word in opinionWords*

 *AND tweet#content NOT CONTAINS 'http' OR 'www.'*

 *THEN WRITE tweet TO file(name);*

 *END*

*END*

In addition, in the first experiment, tweets are filtered by Trade Marks of companies. For example, Apple has lots of registered Trade Marks like iPhone, MacBook and iPad. In order to filter out tweets related to a certain company. I applied filters on all those key words. Most all related tweets are extracted which seems not right. After sentiment analyzing using LingPipe, the sentiment result shows that the three classes records are even distributed on bullish and bearish market days. This proves that the whole set data could not reflect stock market movement.

4.1.2 Sentiment Analysis Tools

Several different sentiment analysis tools and libraries are quite popular.

LingPipe is tool kit for processing text using computational linguistics from Alias-i. Alias-i is an application vendor who provides NLP consulting services. Here I used its sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" often with an implicit category of "neutral". The tool can do sentence based sentiment with a logistic regression classifier. The author also suggests two convinced training dataset. Lillian Lee and Bo Pang have provided annotated slices of movie review data for polarity (both boolean and scalar), and subjectivity. For example, Subjectivity contains 5000 "objective", 5000 "subjective" sentences. Sentences are from Internet Movie Database (IMDB) plot summaries, subjective from Rotten Tomatoes customer review "snippets". The LingPipe provides Java library. First I create a singleton classifier class. Class constructor will take training dataset mentioned above to create an instance of classifier. Singleton is a good practical approach here since training should be executed only once. Then a public classification method is ready to take in sentence. Return will be one of the three sentiment labels. Here provides the pseudo code how I use the classifier in Java Style.

*Class Classifier{*

 *private Classifier classifier;*

 *private Classifier(){*

 *this.classifier = logisticRegression(data);*

 *}*

 *public String classify(String msg){*

 *return sentimentClass;*

 *}*

 *public get(){*

 *if(classifier == null)*

 *Classifier();*

 *return classifier;*

 *}*

*}*

The Natural Language Processing Group at Stanford University is a team working on algorithms that allow computers to process and understand human languages. They provide a NLP processing tool that cover variety of areas. This NLP tool is also integrated in the system as replaceable classifier. By default I use LingPipe as it provides much faster executing speed.

4.1.3 Sentiment Features Calculation

Selected subset now has a sentiment label of positive or negative on every tweet record. Five features are calculated as follows. is calculated as the quantity of daily positive tweets. is calculated as the quantity of daily negative tweets. Bullishness are then calculated by:

 [1]

Bullishness indicates social sentiment on the probability of the stock price will go up. If is larger than , will be a positive number and could up to positive infinity. If is equal to , will be zero. This shows a balance sentiment of bullish and bearish. If is less than , will be a negative number and down to negative infinity. In experiment result, is range from -3 to 2. Message Volume is the logarithm sum of tweets by day.

 [2]

In experiment result, varies a lot, it could be from 2 to 400 depends on which stock or indices is chosen. Most volume data is within a reasonable range, which can ensure its value in the feature set. The fifth feature is Agreement between positive and negative.

 [3]

If all tweet messages about the chosen stock or indices are positive or negative, the agreement will be 1 or -1. If opinions are balanced on each side, will be 0. For almost all the cases, negative sentiment is greater than the positive sentiment; this result in most is range from -0.2 to -0.4. As described by Harvard Business Review's Teresa Amabile, the integral truth behind hypercriticism is that people assume negative statements to be more intelligent than positive ones. This idea has been proven valid over three decades' worth of testing by researchers at both Harvard and Carnegie Mellon [21]. Silent day which no tweet on chosen stock or indices is zero in my result. Only a few cases are with no positive sentiment. This result is significant less than previous research [10] [11]. This benefit should owe to the large dataset I used.

4.2 Finance Features

Finance data of stock price and market indices is collected from Yahoo Finance. The raw data contains daily records of Opening, Closing, High, Low prices and trading volume. Yahoo Finance API allows user collect history data from 1980 to the most recent market day. In this project, I chose the time period that matched twitter dataset, which is from July 31 2009 to December 31 2009. And only market opening days should be chosen, the final number of days is 102 of my allover feature set.

Four finance features are generated from finance data. Closing price is the most important value of daily stock price movement, and the logarithm of is chosen as the first feature . Returns are calculated as the difference of logarithm between the closing values of the stock price of a particular day and the previous day.

 [4]

Returns measure the rise of stock price. Trade Volume is too large and then transformed by:

 [5]

Daily volatility based on multiple intra-day values is calculated using Graman and Klass volatility measures [1-12] with the following formula:

 [6]

The implied volatility of an option is usually compared against historical volatility to see if it is cheap or not. However, while there is only one implied volatility. There are many different measures of historical volatility, which can use some or all of the open, high, low and close. Generally, for small sample sizes the Yang-Zhang measure is best overall, and for large sample sizes the standard close-to-close measure is best.

4.3 Search Engine Query Feature

Preview researches have shown that the search volume can be treaded as a public mood indicator for financial market [12] [13] [14] [15]. In [13], the researcher found out that if negative terms and phrases rise at a time, people present more pessimistic about the economy and market at the same time. To create search query feature, weekly search volumes of stocks and indices are collected from Google Insights for Search (GIS). GIS is a Google web service providing search volume data from January 2004 to present. But only weekly data is available to download, so this feature cannot be added to daily features set. GIS also provide categories for user to focus on specific areas. For this project, Investment and Finance are chosen since they are related to stock market. Examples of 5 companies search volumes are in figure 4.



Figure (4): Search volume trend in history

It shows that the categories are sub divisible, investment and finance are economic related categories but the search volumes are quite different in curves. Volume data is scaled already in a good range from 1 to 100 from Sunday to Monday as a unit.

4.4 Investor Intelligence Feature

Doing an investor survey is the most common approach to collect their sentiment about the market. Investor Intelligence (II) is published by several investor sentiment services. Out of all, in this project I chose the American Association of Individual Investors (AAII) [16]. Their II measures the percentage of individual investors who are bullish, bearish, and neutral on the stock market for the next six months. Individuals are polled from the ranks of the AAII membership on a weekly basis. Only one vote per member is accepted in each weekly voting period.

So AAII do internal statistic calculation based on their member survey votes and generate a weekly feature set of bullish, neutral and bearish which have sum of 100 percent. Two additional artificial features are also included; the bullish 8 weeks average and spread which is bullish minus bearish.

4.5 Correlations between Sentiment features and Finance Features

Five Daily sentiment features and four finance features within the same time period are now ready for calculation. The first steps, the correlations between each sentiment feature with each finance feature are calculated using Pearson Correlation. In statistics, the Pearson product-moment correlation coefficient is a measure of the linear correlation between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation [17]. Pearson-R value great than 0.5 is convinced to are positive correlation and value less than -0.5 is convinced to have negative correlation. In this step, the statistic value can tell if there is any sentiment feature can correlate with any of the finance feature. Later results show that, if the two features have a higher absolute Pearson-R value, it’s more likely to train a better model in the Machine Learning step.

Since the final goal of this project is implementing a prediction model, one hypothesis is the public twitter sentiment will affect near future stock prices. I also did an experiment using one-day lag on stock prices, using sentiment feature of day t to correlation with finance feature of day t+1. There’s also another hypothesis that sentiment cannot predict stock price but twitter sentiment present public opinions and moods of the past stock price changes. In order to find out the true cause and effect relationship, time lag on sentiment experiment is also needed.

GIS and AAII data are weekly based, so the weekly correlations are calculated separately. New sentiment and finance feature sets should align with real market weekly period, this need detailed and carefully coded implementation. There would be holidays when the stock market won’t open, and it result in only four days opening days in certain week. For finance features, open will be the first day open price, close should be the close price of the last day. High and low will be the weekly high and low prices. Trade volumes are summed up and volatility will then be calculated with these new corresponding values. Sentiment features are easier to handle. Quantities of positive and negative tweets are now counted by week.

Weekly correlations are calculated in several different sets. First, the same daily approach gave some results to compare with daily results. Second, GIS-F and GIS-I are features that are proved to have correlations with stock price. AAII sentiment survey provides five new features. Correlations can give some evidences if they are useful in prediction model.

4.6 Machine Learning Algorithms

A list of Machine Learning algorithms is applied on the dataset. Algorithms of different types are selected to find out the best model to use later as the prediction model.

4.6.1 Decision Stump

A decision stump is a model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules. For continuous features, usually, some threshold feature value is selected, and the stump contains two leaves — for values below and above the threshold. However, rarely, multiple thresholds may be chosen and the stump therefore contains three or more leaves. Decision stumps are often used as components (called "weak learners" or "base learners") in machine learning ensemble techniques such as bagging and boosting. Although Decision Stump is not a complicated model, it did can provide good result above average.

4.6.2 Bootstrap Aggregating

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. In this paper, a usual decision tree implementation is used.

4.6.3 Linear Regression

Multiple Linear Regression is a generalization of linear regression by considering more than one independent variable, and a specific case of general linear models formed by restricting the number of dependent variables to one. The basic model for linear regression is

 i = 1…n [7]

In the formula above we consider n observations of one dependent variable and p independent variables. Thus, is the observation of the dependent variable, is observation of the independent variable, j = 1, 2, ..., p. The values represent parameters to be estimated, and is the independent identically distributed normal error.

4.6.4 Gaussian Regression

Gaussian process is a stochastic process whose realizations consist of random values associated with every point in a range of times (or of space) such that each such random variable has a normal distribution. , t ∈ T, for which any finite linear combination of samples has a joint Gaussian distribution. More accurately, any linear functional applied to the sample function will give a normally distributed result. Notation-wise, one can write X ~ GP(m,K), meaning the random function X is distributed as a GP with mean function m and covariance function K.

Inside Kernel used:

 [8]

4.6.5 Radial Basis Function Network

The radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters.

RBF networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer.

Figure (5): Network Structure

An input vector x is used as input to all radial basis functions, each with different parameters. The output of the network is a linear combination of the outputs from radial basis functions.

4.6.6 Multilayer Perceptron

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network.

4.7 Present Day/Week Record Generation

A prediction system is the ultimate goal of this project and final step. Machine learning results get from last step show that each stock may need different models to get the best correlations and accuracy. For chosen stock, models are already generated and applied for new input data. If new stock is needed, another repeat of previews steps are required.

Twitter Streaming API provides real time data and it’s easy to set up. Listeners for wanted stock and companies are added for getting separated intermediate files. Tweets of each file also go through filtering and cleaning component. The filtered out tweets then go into sentiment classifier. Daily and weekly Counters of each stock stores positive and negative tweet quantity for calculating sentiment features. Finance features can be collected easily from stock market API or service providers like Yahoo Finance. Search Engine Data can be collected from Google trend on daily base. That’s more data since it only provide weekly data in history. AAII present sentiment survey are not free, membership are required to access new results, so it’s not part of current prediction system.

When stock market close, feature sets are extracted from raw data and put into predict model to get prediction for the next day’s stock prices.

5. Results and Analysis

5.1 Sentiment Feature set

11 leading technological company and two market indices are chosen in this project. In order to get sentiment feature set for each target, several steps are need. Here I use Apple as an example to show the intermediate result.

As described before, the raw twitter data contains 400 million tweets from overall the world in different languages. Opinion filter and language filter are applied to extract target related tweets that can indicate stock prices movement. 42 thousands tweets are filter out for Apple for five month. This is a very restricted subset of all Apple related data. Previews result for all Apple related tweet is 3 million, but contains lots of usual life moods and market advisements. How to filter on tweets that correlation with stock market is a very important part of building up a good model. A group of researches choose the opinion filter like what I did here.

A sentiment classifier called JLangDetect with preview training classifier model was used to classify a tweet into 3 classes, positive, negative and neutral. I do focus on positive and negative classes and the neutral part is removed. Positive and negative quantities are counted by day. An example on day record of sentiment feature set will have Date, Positive Quantity, Negative Quantity, Bullishness, Message Volume, and Agreement like in figure 5.



 Figure (5): Example of Sentiment Feature Set

Sentiment feature set has a time series of daily records from July 31 2009 to December 31 2009. After alignment with stock market opening days, a set will have 102 records. An example of features trend is in figure 6.

Figure (6): Example of Sentiment Features Trend

5.2 Finance Feature set

Finance figures can be easily collected from web service like Yahoo Finance. Raw data contains Date, Open, High, Low, Close, Trade Volume. Four finance features are then calculated; Return, Close, Trade Volume, Volatility. Figure 7 is an example.



Figure (7): Example of Finance Feature Set

5.3 Weekly Feature set

Weekly feature set contains the same feature sets above with new features get from GIS and AAII. Daily features are processed into weekly features. GIS provides two additional features Search Query in categories of ‘Finance’ and ‘Investment’. AAII provides investor survey data, which result in five new features bullish, neutral, bearish, spread, and 8-week average bullish.



Figure (8): Example of Weekly Feature Set

Figure 8 shows the new seven weekly features in period of one week.

5.4 Correlations and Analysis

I measure the correlations between finance features with all other features using Pearson Correlation. The closer to 1 or -1, the better the correlations are.

5.4.1 Daily Correlations

Daily correlations can be divided into four parts by the four finance features. Experiment results show that Return and Volatility only have weak correlations with sentiment features. Return has average Pearson-R of 0.1 and Volatility has average Pearson-R of 0.2. Strong correlations are found on Close.



Figure (9): Correlations between sentiment features and Close

In figure 9, First 6 sets of Positive, Negative, Message Volumes have Pearson-R value less than -0.5, which means strong negative correlations. The strongest is Amazon TV with value of -0.814. For DJIA, TV shows positive correlation 0.65 even Pos and Neg have low value. DJIA is the only case with accepted correlation of Bullish. This could because of the most relevant tweets are filtered out. People talking about Dow Jones Indices with concerns of stock market only while other topics on companies may vary a lot.

5.4.2 Weekly Correlations

In weekly Correlations, two important additional features are added into feature set. Search Engine Data collected from Google Insights Search provides indicators of stocks in two categories, Investment and Finance. After processing the daily records into weekly records and combining with GIS data, correlations are given in figure 10.



Figure (10): Weekly Correlations between sentiment features and Close

Statistically, absolute average correlation values increase from 0.417 to 0.572 for selected stocks shown above. This improvement shows weekly sentiment of twitter can better indicate stock price movement. GIS features didn’t correlate well as others features, but it did reach as high as 0.925 in DJIA case. For all GIS features, absolute average value of 0.46 could prove them as valuable features. Comparing GIS-F and GIS-I, there are cases like Amazon and DJIA, which both of them perform well. There are also other cases one of them performs better. I would keep both of these two features for the predicting models so far.

AAII provides investor survey data which present investor sentiment on up coming market week. Since AAII required membership to access their full database, I was restricted to survey data on S&P 500. S&P is also a market indices, so I only chose NASDAQ and DJIA for further calculation. Correlations of five AAII features with finance features are given in figure 11.



Figure (11): Weekly Correlations on AAII feature set

Very Strong correlations can be observed from this table on Close and Trade Volume. The highest is 0.9956 on bullish 8 weeks average with DJIA Close and 0.9968 on DJIA Trade Volume. So far, the AAII provide strongest indicative feature set.

5.4.3 Correlations with Time Lag

In methodology, I have two hypotheses. H1 is twitter sentiment can predict the stock price of near future. H2 is twitter sentiment present public mood on stock price of the same day or the day before. In order to find out the truth, experiments with time lag on each feature gave the following results.



Figure (12): Correlations between sentiment features and Close

Figure 12 show different time lags of Close correlating with Message Volume. Obviously, one-week lag of Close got the highest correlation than other time lag. This proved Hypothesis one; sentiment features can predict the stock price of the following week.

5.5 Machine Learning and Models

Six different Machine Learning algorithms are applied on the feature set described above to get prediction models. Algorithms vary from Decision Tree, Regression Model and Neural Network. Based on 4.4 conclusions, one-week lag is applied on finance features.

Daily feature set contains 5 sentiment features and Close is chosen as the target.



Figure (13): Machine Learning Results

In figure 13, algorithms appear to have equal performance on each target. Comparatively, Gaussian Process is slightly better than other algorithms and Multilayer Perceptron is the worst. Correlation values of models reflect exactly what I have in preview step. An unexpected case is DJIA figures. It got signification lower value than the feature correlation value. Figure 14 gives comparisons between daily and weekly models.



Figure (14): ML daily & weekly comparison

Value sets with ‘–w’ are weekly sets. It’s clearly that weekly-based models have better correlations than daily-based model. Statistically, average absolute value increased 0.1.

Weekly data set also added AAII data, which gave strong correlations in 4.4.2. AAII can only indicate marked indices so I did experiments on NASDAQ and DJIA. In order to compare performance improvement with new features, I did experiments in four sets; origin sets without new features, adding AAII features, adding GIS features, adding AAII and GIS features.



Figure (15): ML GIS & AAII comparison

In figure 15, DJIA results show that Gaussian Process is the best model across 6 models providing the highest average correlation. But without additional features, origin sentiment feature gave the best performance using Multilayer Perceptron model. Combining all feature sets didn’t always improve improved model performance. Instead, a particular model and features set needs carefully chosen to give the best performance. This result shows no model can replace others and multimodal combinations are required in predicting system.

NASDAQ histograms are more stable, results show origin sentiment features gave better correlations than others.

All the results are generated using cross validation. The fitting process optimizes the model parameters to make the model fit the training data as well as possible. If we then take an independent sample of validation data from the same population as the training data, it will generally turn out that the model does not fit the validation data as well as it fits the training data. This is called overfitting, and is particularly likely to happen when the size of the training data set is small, or when the number of parameters in the model is large. Cross-validation is a way to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available [18]. Overfitting is quite possible in this project since training data is limited.

In order to verify if the model is overfitting or not, another set of experiments using randomly split training (90%) and test (10%) on daily dataset gives the following results.



Figure (16): ML training 90%, test 10%

Average correlation value of all experiments increase from 0.568 to 0.632. Decision Stump shows equal performance while RBF Network decrease on all stocks. Other four algorithms have a little bit higher correlation value than cross validations. These results proves the cross validation can prevent predicting models from overfitting and the previous results are valid.

 

Figure (17): Amazon predicted value comparing actual value

Figure 17 gives an example of predicted value comparing with actual value. The predicted values clearly align with actual value at the beginning 1.5 months and the ending 2 month. The overall correlation is quite high. The weak part of this prediction is the error. Based on observation, the maximum error is up to 0.12, about 2.5%. If the is restored to close price, the actual prices error is 15. It’s 12% error, which should be consider a bigger error.

5.6 Real Time Twitter and Daily Data Collection

The prediction system consists four components, External APIs, filtering and cleaning, new feature generation and prediction model. New data are collected from three external APIs. Twitter Streaming API will keep pull real time tweets by stock and indices.



Figure (18): Twitter Streaming data

Figure 18 gives a part of streaming process logging. A log includes total tweets received, date, and contents of sample tweets. Simple format tweets contains the only information needed for this project, data and content. The procedure follows this pseudo code.

*Function GetRealTimeTweets*

*establishConnection();*

*query(names);*

*FOR each marketDay*

 *FOR each tweet IN queue*

 *tweet = queue.take();*

 *parse(tweet);*

 *IF tweet#language eqauls 'EN'*

 *AND tweet#content CONTAINS word in opinionWords*

 *AND tweet#content NOT CONTAINS 'http' OR 'www.'*

 *WRITE tweet TO file(name);*

 *END*

*END*

Real time tweets from twitter are in JSON format, details can be found on Twitter Developers [19]. The original tweets contains tons of tags, flags and coordinates. These additional information can provide more details about a tweet and now under wide research. In this project, I implemented a simple format parser that can exact date and content. Tweets also go through further processes filtering.

Yahoo Finance API is quite similar as Twitter’s. Registration of applicaion on its website is required. Afterwards, the YQL (YahooQuery Language) platform enables you to query, filter, and combine data across the web through a single interface [20]. It exposes a SQL-like syntax that is both familiar to developers and expressive enough for getting the right data. JSON and CSV format are two output options but also of them need parse for next step.

Google Search Volume can be collected only by hand since I need to specifiy the query.

6. Future Work

There are three parts of this work that can have improvement, filtering component, sentiment classifier and machine learning algorithms.

Filtering component now consists of stock name filter, language filter, opinion filter and spam filter. These are standard filters developed from early research. After the filtering process, gigabytes of data reduced to megabytes, it’s every like valuable tweets are filter out. Besides these static filters, more dynamic filtering method should detect related tweets with the help of social network. Twitter APIs provide friends list and number of likes. Some people more influential could affect people in his/her friends list and then spread through the network. This information flow could provide a way to collected more stock market related tweets and public sentiment.

The sentiment classifier component in this work uses a manually labeled movie sentiment dataset as training data. In experiment results, around half of stocks and indices always have a larger number of negative tweets than positive tweets. This results in mostly negative of bullish, which cannot reflect the return value. This result could be quite different from what was expected in stock market related tweets. The next step, I want to build a training dataset based on tweets sentiment of stocks. I hope the sentiment classifier can have better performance from the new training data.

When doing predicting models, I used six existing algorithms with default parameters. Although strong correlations are generated using these algorithms, the efficiency of them need long term testing. Most importantly, the mean absolute percentage error is comparatively high. My result and previous work have MAPE no better that 1.5%. And this result is base on logarithms of stock price. By restoring the original value, the different of predicting value and actual value could have more than 10% error. Better machine learning algorithms are desired.

7. Summary

The thesis describes a predicting system for target stocks and indices movement combining twitter, search engine and investor intelligence data. Sentiment features are generated from twitter dataset using sentiment analysis. Strong correlations are found on sentiment features with return feature for both daily and weekly dataset. Time lag results show sentiment features of one week can best predicting the return value of the following week. Combining sentiment features with GIS and II feature set, the predicting system achieve up to 0.8 correlation value. Finally, by adding real time APIs and processing components, the prediction system can take in present day and week records to generate predicting values for the next day and week. Overall, this system is working with reasonable efficiency and the built-in methods show strong evidence of correctness.

8. References

1. Stock Market Capitalization By Country. <https://www.quandl.com/c/economics/stock-market-capitalization-by-country>
2. How does the stock market affect the economy?

<http://www.economicshelp.org/blog/221/stock-market/how-does-the-stock-market-effect-the-economy-2/>

1. Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. arXiv.org, arXiv:0911.1583v0911 [cs.CY] 0919 Nov 2009.
2. H. Mao, S. Counts, and J. Bollen, “Predicting financial markets: Comparing survey, news, twitter and search engine data,” arXiv.org, Quantitative Finance Papers 1112.1051, Dec. 2011
3. Rao, T., and Srivastava, S. Analyzing stock market movements using twitter sentiment analysis. In Proc. of ASONAM (2012).
4. J. Yang, J. Leskovec. Temporal Variation in Online Media. ACM International Conference on Web Search and Data Mining (WSDM '11), 2011.
5. J. Bollen, H. Mao, and X.-J. Zeng, “Twitter mood predicts the stock market,” Computer, vol. 1010, no. 3003v1, pp. 1–8, 2010.
6. JLangDectect language detector. <https://github.com/melix/jlangdetect>
7. JLangDetect 0.3 released with bundled language profiles.

 <http://www.jroller.com/melix/entry/jlangdetect_0_3_released_with>

1. M. Z. Frank and W. Antweiler, “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards,” SSRN eLibrary, 2001.
2. T. O. Sprenger and I. M. Welpe, “Tweets and Trades: The Information Content of Stock Microblogs,” SSRN eLibrary, 2010.Affectiva, Q Sensor 2.0 Datasheet, 2012, http://www.affectiva.com/download/Affectiva\_Q\_User\_Manual.pdf.
3. I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen, and I. Weber. Web search queries can predict stock market volumes. http://arxiv.org/abs/1110.4784.
4. Z. Da, J. Engelberand, and P. Gao. The sum of all fears: investor sentiment and asset prices. http://ssrn.com/abstract=1509162, 2010.
5. Z. Da, J. Engelberand, and P. Gao. In search of attention. Journal of Finance, 66:1461–1499, 2011.
6. T. Preis, D. Reith, and H. E. Stanley. Complex dynamics of our economic life on different scales: insights from search engine query data. Philos Transact A Math Phys Eng Sci, 368(1933):5707–19, 2010.
7. AAII Investor Sentiment Survey. <http://www.aaii.com/sentimentsurvey>
8. Pearson Correlation. <http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient>
9. Cross Validation. [http://en.wikipedia.org/wiki/Cross-validation\_(statistics)](http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29)
10. Twitter API, Twitter Developers. <https://dev.twitter.com/overview/api/tweets>
11. Yahoo Query Language. <https://developer.yahoo.com/yql/>
12. Negative Tweets More Likely to Spread Than Positive Ones. <http://bigthink.com/ideafeed/is-negative-content-more-likely-to-go-viral>