Paper based Stock Market Trader on Twitter

NATURAL LANGUAGE PROGRAMMING
FINAL PROJECT

12-1-2017

UNIVERSITY OF TEXAS AT DALLAS

Poonam Purushottam Pathak
Netid: (ppp160130)

ABSTRACT

The overall sentiment built around a company greatly impacts its stock market. Therefore, statistical information from NASDAQ is not enough to do efficient stock trading. What people perceive about a company is also important to the analysis.

The goal of this project is to be able to do paper based (virtual) stock trading based on the sentiment analysis performed on the latest updates on various company stocks from important financial services via social networking site like twitter and news websites like Reuters.

APPROACH

My approach to this problem is to make use of twitter updates combine it with historical data to calculate a relative score for each stock. The data will have to be retrieved via twitter streaming app and crawling on popular new websites like Reuters. For the sentiment analysis, firstly important information such as company name, investment percentage will be extracted from the tweets. The tweets will then be classified under a class based on Maximum Entropy Modeling method.

The sentiment derived from tweets and news website and its current position in the marketplace, with a higher differential will be used to calculate the final score for each stock. Top ten stocks with higher score will be then recommended as part of stock trading.

1) Stocks Market Domain Knowledge

Before deep diving in to the project details, following are the basic terminologies used in

1.1 Call Options

A call option is an agreement between two investors that gives one the option, but not the obligation, to buy a stock of a company at a set price, known as the strike price within a certain time- frame. This is when the investor will get the sentiment that a company’s stock price will increase in the future. Based on this sentiment the investor buys the stock at the current price, eventually seling the stock if the company’s stock price has increased.

If, however, a company’s stock value has de- creased during this time, then there is no reason for the investor to buy at the strike price, when buying on the open market is cheaper. In this case, the investor will have lost the amount paid as a premium for the right to buy the stock at the strike price.

For example, if an investor bought a call option for 500 shares of IBM Stock with a strike price at
$100 when the share price is at $90, set to expire in 2 months, the investor would theoretically pay
$5000 as their premium cost ($10 x 500 shares). If the stock price increased to $100, the investor will still be at a loss for $5000 due to the premium cost. The investor will begin to make a profit when the IBM stock price exceeds $110, when the $10 gross profit covers the $5000 premium cost.

1.2 **Algorithmic for Strike Price Determination**

The algorithm designed for strike price determination is as follows:

1) The algorithm divides each option into two categories: ”Soft” and ”Hard”.
2) Soft option generated in the trade file whenever the general sentiment around a company affects its Vi- ability Score (VS) to the point of passing the soft threshold, which in our program which is initially set to +/-200 VS.
3) Hard options (hard calls and hard puts) are generated in the trade file whenever the general sentiment around a company affects its VS to the point of passing the hard threshold, which was set to +/- 500 VS.

As we can see the that if the general sentiment surrounding a company (taken from the tweets and news updates from websites like Reuters) becomes negative, then the company’s corresponding VS will drop extremely quickly as well.
On the other hand, assuming that there’s a general positive sentiment trend for a company, its VS score will gradually increase enough to the point where a soft call will be generated instead.
Thus including the sentiment analysis for a company from twitter and news sites is extremely important for doing the stock trading of companies.

2) **Training the Sentiment Analyzer:**

2.1 Training Dataset

- The dataset was first created based on set of defined positive and negative emoticons such as :) and :(.
- The tweets collected are basically classified based on the emoticons used by the users (The reality of sarcastic emoticons to describe mixed emotions is not taken in to the account for this project). The classified tweets are then saved in to csv file.
- These classified tweets ate then stripped out of the emoticons and preprocessed as per the following:
  - Lower case - converted the tweets to lower case
  - URLs - eliminated all the URLs
  - #hashtag - hash tags give us some useful information, so replaced them with the exact same word without the hash.
  - Punctuations and additional white spaces - removed punctuation at the starting and ending of the tweets.
  - Feature Reduction (For example: Tokenization, Removing Stopwords, Twitter symbols, and Repeated Letters)
  -
- preprocessing the data causes the classifier to learn from the other features present in the tweet. So that classifier can use the non-emoticon features to determine the sentiment.

2.2 TRAINING THE CLASSIFIER

- Maximum Entropy Likelihood is the classifier used here to create a sentiment analyzer.
- NLTK library is used to train the data and classify based on probability distribution.
- The training data stored in the csv file is first fed to the Maximum entropy classifier of NLTK.

- To test the accuracy, it was then run against the test data which included 25 percent of the mixed tweets from training set.
- The accuracy currently came out to be 93.57 percent on training data and 80.45 percent on test data.

3) Fetching Stock Market data:

Major stock market data is fetched from the following:

1) Tickers list form Nasdaq
2) Fetching the latest news (not older than 40 days) from Reuters website using python library BeautifulSoap4 and saving it in the form of csv.
3) Running the front end engine with help of Python library Tweepy , to retrieve a batch of 3000 tweets from the following major tending user accounts popular for stock market monitoring:

```
"1364930179"   # Warren Buffett
"14886375",    # Stock Tweets
"15897179",     # breakoutstocks
"28571999",       # bespokeinvest
"2837841",      # CNNMoneyInvest
"16228398",     # Mark Cuban
"1754641",     # nytimesbusiness
"21323268",     # NYSE
"184020744",    # Mike Flache
"19546277",     # YahooFinance
"778670441405775872", # MarketsInsider
```

Algorithm to calculate Viability Score

1) At first we initialize the VS score for the companies listed out in tickers list under History.csv file with 200 points as average.

2) For each tweet, we do a first pass to extract out any company names that can be found in the tweet, based off of a pre-defined list of stock tickers, full company names, and a selection of colloquial names of companies.

3) A Tweet tweet is represented in a tuple, where:
   tweet[0] -> Tweet message
   tweet[1] -> Number of retweets, basically our measure of how important/wide-spread this tweet is
 The number of retweets will act as the weight of that tweet to impact the sentiment of that company.

4) If the extracted company in the tweet matches with the company listed in the tickers list, the the tweet message is passed to the MEM classifier we built and stored. This classifier will output the sentiment polarity value ranging between 0 to 1.

5) 1 stating positive sentiment and 0 stating the negative sentiment.

6) The same way the news messages are extracted from csv list for Reuters and their sentiment polarity is determined.

7) After we get the sentiment polarity from the tweets and reuters , These values are used to update the VS score as follows:

```
total_score = viability_scores[new_company][0] + (new_score * new_weight)
total_weight = viability_scores[new_company][1] + new_weight
```

Scope of improvement :

The training set created to train the sentiment analyzer can be improved by exploring more than just the classification based on the emoticons. In reality the emoticons used by the users may not convey the straightforward emotion or sentiment as we see theoretically. This can significantly impact the training.

Conclusion

By analyzing the sentiment around Twitter users, it is possible to generate a ”sentiment value” score about publicly traded companies, and to then use these trades in order to generate a profit that could beat the market in the short term. The current pa- per profit yield of 1.8% while the S&P index had a growth of 0.02% showcases the ability to lever- age sentiment analysis in the pursuit of intra-day profits