Correlation between Twitter Sentiments & Company Stock Return

Submitted by
NIKHIL LOHIYA
TARANPREET SINGH
KUMAR BIPULESH
VIGNESH SRIRAM
FERNANDA CAPELA

Mentor
PROF RONG LIU

WEB ANALYTICS (BIA-660-C)
Fall 2017

STEVENS INSTITUTE OF TECHNOLOGY.

INTRODUCTION & MOTIVATION

Volatility is one of the most complex characteristics of the stock market since it brings a great amount of risk to investors. Yet, it's also what opens opportunity (arbitrage) for profitable investment for the ones that understand how it works. There are several factors that influence stock market volatility. But one of the hardest to understand and anticipate is human emotion.

On a different end, micro-blogging has become a popular communication tool, and for that can aggregate opinions of a lot of people with the different background in any place and time. This information has been used in both academic and industry to obtain precious feedback from massive crowds with a lower cost and faster than pace than surveys, for example. Therefore, sentiment analysis is becoming a powerful decision-making tool nowadays. Before investing in a company, one can leverage the opinions of the people about that company to find out where it stands.

For years financial stock market predictors have yearned to know the future market and with the advent of machine learning technologies, their job has become quite simple. When we join hands of technology such as twitter to a domain such as Financial Market analysis we see different new concepts such as Behavioral Finance, which is the reason, we started this project. Our project wants to explore the relationship between people's impressions towards certain companies and their stock market variations.

OBJECTIVE

To find out sentiments of tweets associated to a company, and relate the sentiments to the market return of the company.

RELATED WORKS

A number of prior works are designed to predict stock price using Twitter or another social network. One among them is by Johan Bollen, Huina Mao, Xiao-Jun Zeng in which they attempt to predict the data of the tweet by classifying it into 6 different categories (Calm, Alert, Sure, Vital, Kind, and Happy). Using fuzzy logic, they got an accuracy of 87.6%. They analyzed the text content of daily Twitter feeds by two mood tracking tools, namely Opinion Finder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood which helps them to categorize the tweet in above mentioned types.

METHODOLOGY

Data

Since our goal is to compare variations between market price and people's opinions, we had to gather data from both stock market and Twitter. The financial data doesn't need a lot of manipulation, since we just want to obtain the movement flow on a specific period of time. The twitter data, however, needed special attention.

First, we collected tweets using keywords related to the companies we wanted to evaluate. Then we performed the data preparation process, starting with data cleaning, tokenization and normalization.

For our **training set** of tweets, we have used the following benchmark:

- 4 – POSITIVE TWEET
- 2 – NEUTRAL TWEET
- 0 – NEGATIVE TWEET

Previously we planned to use open source labelled dataset (provided by Stanford) tagged for about 160k tweets but it was giving poor results when we ran with our algorithms. Moreover, in real benchmarking tweets are either Positive/Negative/Neutral. We manually tagged 2500 tweets for our analysis of only 2 stocks i.e. AAPL and MSFT.

For converting the value from the above stated metric system to Binary (i.e. 0 for negative mood and 1 for positive mood), we performed following steps:

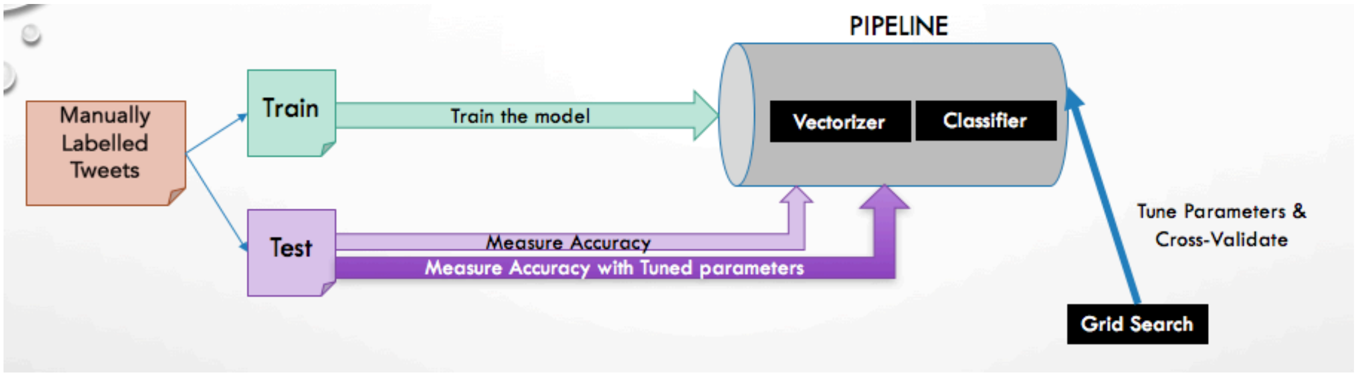
- Take average of all the Sentiment numerical value of all the tweets in a day.
- If the average value of the tweets is above 2 then the value is assigned as True or 1.
- If the average value is less than 2 then the value assigned is False or 0.

In the Stock Market Data, we have taken the daily returns by calculating the difference between the closing Price of One date and the Closing price of the previous day. (This takes into account the pre-trading hour values as well.) The values are then converted to Boolean between 0 and 1 by assigning Positive returns as True or 1 and the negative returns as False or 2.

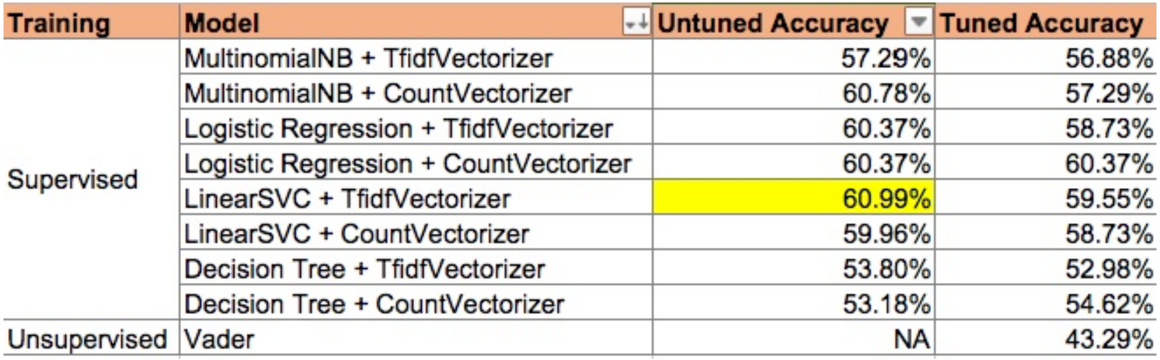
Once the Boolean values from both the sentiments and the stock market is received a pandas data frame is created by using the **inner join** function to remove the redundant values (Weekends and Public holidays). The correlations are then calculated with the values such that the same values are returned as true in the final output.

Classifier Selection

Note: ***For detailed steps of the Classifier Selection, kindly refer the Jupyter Notebook "Classifier Selection.ipynb".***



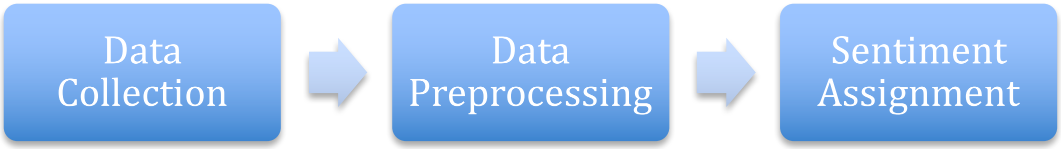
Classifiers comparison:



Training	Model	Untuned Accuracy	Tuned Accuracy
Supervised	MultinomialNB + TfidfVectorizer	57.29%	56.88%
	MultinomialNB + CountVectorizer	60.78%	57.29%
	Logistic Regression + TfidfVectorizer	60.37%	58.73%
	Logistic Regression + CountVectorizer	60.37%	60.37%
	LinearSVC + TfidfVectorizer	60.99%	59.55%
	LinearSVC + CountVectorizer	59.96%	58.73%
	Decision Tree + TfidfVectorizer	53.80%	52.98%
	Decision Tree + CountVectorizer	53.18%	54.62%
Unsupervised	Vader	NA	43.29%

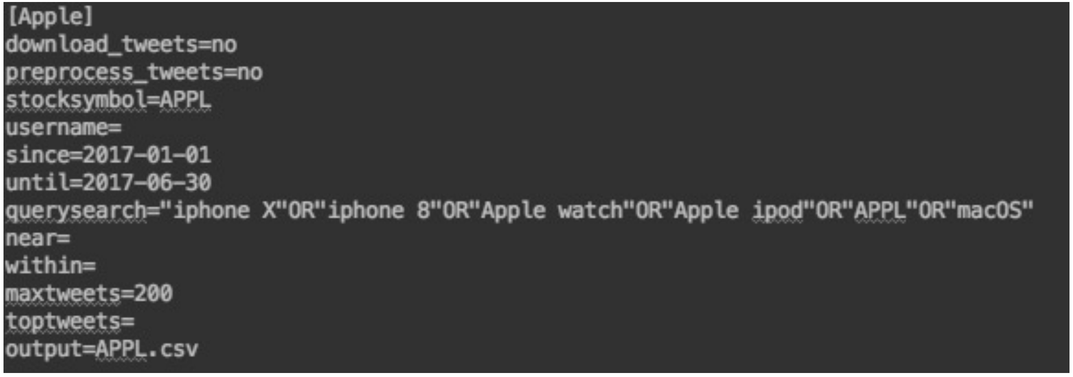
Twitter Data Collection & Sentiment Analysis

Note: Code for this part is placed in the folder named "TwitterSentimentAnalsis"



<u>Code Flow</u>

Step 1: Customize 'configuration.ini' for the Twitter Data Collection



```
[Apple]
download_tweets=no
preprocess_tweets=no
stocksymbol=APPL
username=
since=2017-01-01
until=2017-06-30
querysearch="iphone X"OR"iphone 8"OR"Apple watch"OR"Apple ipod"OR"APPL"OR"macOS"
near=
within=
maxtweets=200
toptweets=
output=APPL.csv
```

- New section as above marked by company identifier for additional companies. For example, [Apple], [Google] etc
- Control downloads and preprocessing by the 'download_tweets' and 'preprocess_tweets' flag
- Changes to 'since' and 'until' parameters to define the time period for tweets collection
- Change 'querysearch' for the search parameters
- Change 'maxtweets' for the maximum number of tweets to be downloaded for each day

Step 2: Trigger Data collection and Preprocessing

- Command: "python driver.py configuration.ini"
- Raw data is stored in the folder "data/twitter_raw_data" (if the 'download_tweets' is set to 'yes' in configuration.ini)
- Once raw data is collected data is preprocessed/cleaned and stored in "data/twitter_clean_data" (if the 'preprocess_tweets' flag is set to 'yes' in configuration.ini)

Step 3: Sentiment Analysis

- Execute 'sentiment_supervised.py' after changing *input_file_name* variable value (line 39) using command: "python sentiment_supervised.py".
- Final file with Tweet sentiments is stored in "results/sentiment_analysis_LinearSVC" folder.

Stocks Data Collection & Conversion

***Note:* For detailed steps of the Classifier Selection, kindly refer the Jupyter Notebook "StockData-Plotting-Notebook.ipynb".**

Code Flow

Step 1: First we fetch the data from 'Yahoo' finance api using Pandas Datareader library.

```
Data.DataReader("AAPL",'yahoo',dt.datetime(2017,1,1),dt.datetime(2017,6,30))[['Close','Open']]
# input format yyyy/mm/dd
```

Step 2: We get the 'Open' and 'Close' balances and find the return using the below formula

```
Data1['return'] = (data1['Close']-data1['Open'])/data1['Open']
```

Step 3: Next, we take in the predicted file for that company and group these values by date

Step 4: We compute the mean for each day and reduce the value by 0.5 (this is done to overcome the problems of neutral data)

Step 5: We then create another data frame to store the true/false values if the mean for that day is above 2

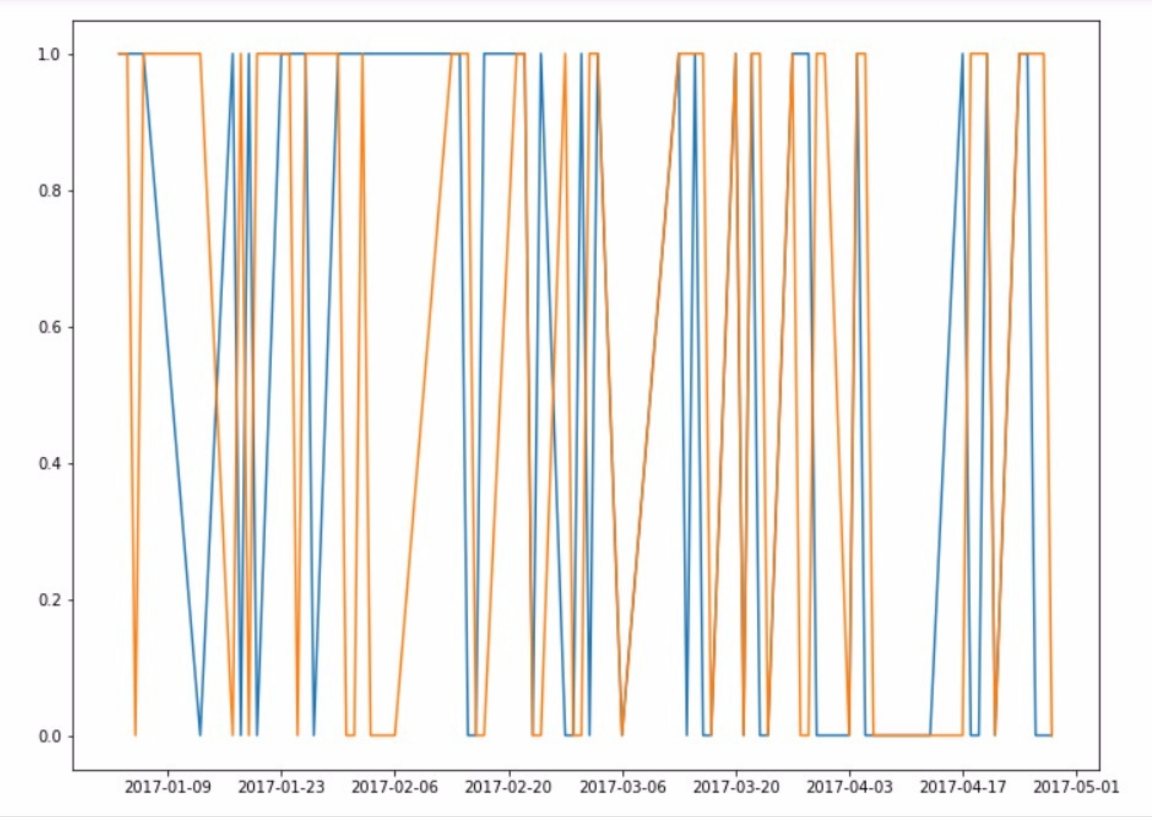
```
avg = by_date.mean() - 0.5;
avg2 = avg>2
```

Step 6: We join both these table using 'inner' join based on the common columns (date) and plot the graph.

EXPERIMENT RESULTS ANALYSIS & CONCLUSION

Using the manually labelled tweets of Apple for the month October 2017, we observed a match of 72% between the day-level sentiment of Twitter to the daily return of Apple stock. We took this as the benchmark for our experiment.

As an outcome of our experiment for Apple (using data from January-2017 till June-2017), below is the pattern of match between Twitter Sentiment and Apple stock movement. We observed 52% match between the daily-level sentiment of Twitter and the stock movement.



Overall the experiment result might lead to the conclusion that there is no direct relation between Twitter Sentiment and Market movement for a company's stock, however, we need to keep the following points into consideration:

- The daily maximum number of tweets that we chose to do this experiment was limited to 200, which might not actually be a true representative of the actual public sentiment for a day.
- Majority of the tweets were labelled as 'Neutral' which further decrease the actual representation of 'Positive' and 'Negative' sentiment.

FUTURE WORK

- In the future, the relationship between the tweets of some important twitter accounts and the whole stock market can be investigated using the generalization of this project.
- Apart from this feature, we can also analyze the tweets(opinions) of important market strategist, experts and senior management of the company of these accounts to test the relationship.
- Currently we are only exploring twitter but we can include more documents like news articles or articles written by stock market experts.
- Increasing the volume of tweets labeled manually, and having a good balance of tweets with positive, negative and neutral sentiments, will help to generate a better TFIDF which in turn can help increase the performance of the model.
- Advanced Techniques such as Neural Networks can be a viable option if the training data size is large enough.
- A Prediction model which takes into consideration more variables like the period of year, company's previous 30-day average return etc. might be helpful to achieve a higher accuracy in relating with Stock price movement.