Twitter Market

By: Ben Walczak

https://github.com/ben-walczak/TwitterMarket

Abstract:

Deborah Dian of Daily Voice News examines how Trump's tweet dropped the Boeing stock by 1 billion dollars. She goes to say, "On Dec. 2, 2016, President-elect Trump surprised investors with a Tweet saying, 'Boeing is building a brand new 747 Air Force One for future presidents, but costs are out of control, more than $4 billion. Cancel order!'[1]… The impact on Boeing stock shocked investors who watched as the company dropped $1 billion in value the morning he released the Tweet." The impact of a tweet can have a tremendous impact on the stock market. With high frequency traders, the stock market, now more than ever, can be easily be manipulated by social media in the matter of minutes. Generally, social media is available to the public. So, one could easily extract tweets from Twitter. The sentiment of tweets throughout a day can then be processed and compared to the stock market prices to determine whether relationships exist. Other metrics of tweets could also potentially have a relationship to the stock market.

Introduction:

Tweets containing the string, 'Nasdaq', will be extracted on Python, using Tweepy. Other modules will assist in text processing and graphs and the setup of sentiment analyzers. Valuable metrics included in extracted tweets are text of tweet, time created, and number of retweets. Tweets will be compared to the Nasdaq composite price.

Research questions considered during the project:

- How does the cumulative sentiment of tweets throughout a day relate to the stock market price?
- How do tweets with more retweets relate to the stock market price?
- How does the sentiment of tweets relate to the highs and the lows of the stock market price throughout the day?
- Does the frequency of tweets relate to how much the stock price varies?

There are three sentiment analyzers to be used. One will judge sentiment based on words within a tweet that often indicate whether a stock is doing good or bad, this being preliminary sentiment analysis. Some of these indicators include buy, invest, up, sell, cut, down, etc. Another sentiment analyzer will be Naive Bayes algorithm, tested and trained on a local sample set. The last sentiment analyzer will be NLTK sentiment analysis, which is a pre-tested and pre-trained sentiment analyzer on an outside sample set.

Processing of Tweets:

Extracting meaning out of text can be a difficult thing. Many factors exist such as slang, sarcasm, etc., but in tweets even more factors are introduced such as hashtags, URL's, retweets, replies, etc.

---

[1] Dian, D. (2017, January 8). How Trump Tweets Affect the Stock Market - Is It Stock Manipulation? Retrieved May 5, 2017, from http://thedailyvoicenews.com/2017/01/08/how-trump-tweets-affect-the-stock-market-is-it-stock-manipulation/

For the sake of simplicity, URL's were removed within each tweet without considering the actual URL. Sentiment analyzers would have a tough time extracting meaning from URL's, and it would be time consuming and difficult to follow up the URL. Any tweet directed at another user or retweeting a user using the @ was removed of this symbol and the string that followed. Users were not considered when extracting tweets, unless the tweet had a significant number of retweets.

Another difficult aspect of tweets to handle is tweets containing updates on the changes in stock prices. To handle this, any tweet containing a '-' symbol or '+' symbol followed by a digit was assumed to refer to stock prices. This method is prone to errors, but it is simple and easy to implement. The symbols '-' and '+' were then removed from each tweet along with its change in price.

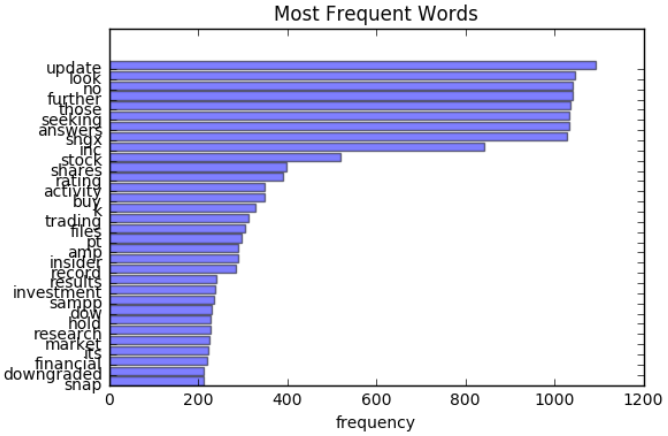
Any hashtag within a tweet was also removed and stored into a variable for later use.

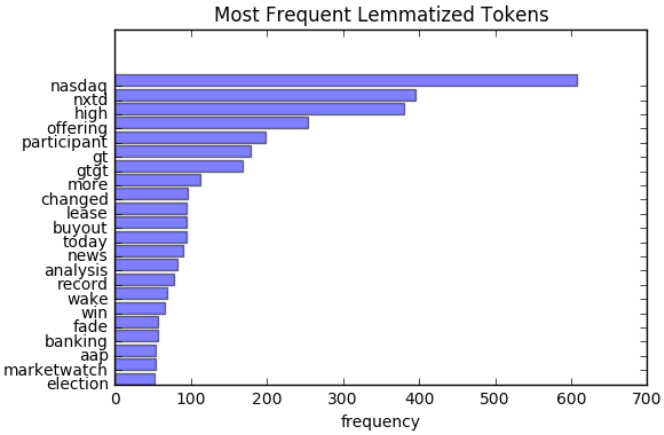
Each tweet was also cleaned of any unnecessary whitespace and converted all to lowercase. Additionally, stop words were removed from each tweet to aid in sentiment analysis.

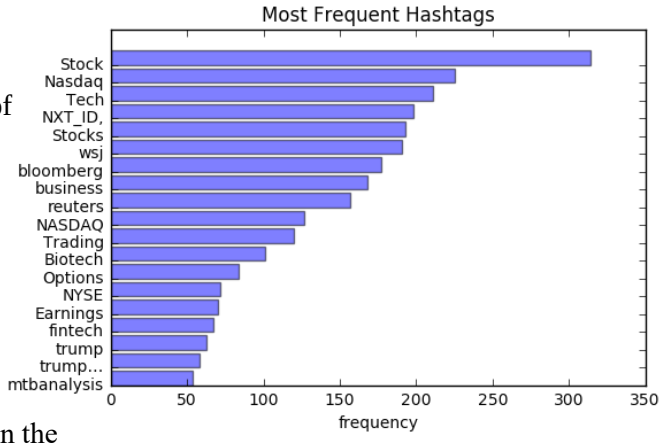
The unfortunate part about text processing is although the altered text is easier to analyze, it is not the same as the original text. What has been created is artificial data. It may be insignificant, but each stopword or string that was removed carried some meaning, even if it may have been small.

Display Frequent Terms:

It's important to take note of frequent terms within tweets, whether it be part of the tweet or the hashtag. Many of the frequent words within tweets were nouns or actions related to stock markets. For example, sampp or better known as S&P500 was mentioned as well as buy. Some words carry little meaning and some word carry significant meaning. Topic modeling would be able to do a much more in depth analysis of words often grouped together, giving new meaning to frequent words.







Since hashtags were removed from tweets it is important to look at frequent unique hashtags, some of which include NXT_ID, bloomberg, Trump, and mtbanalysis. As mentione in the abstract, the president can carry significant influence in the stock market when it comes to tweeting. A close monitoring of Trump's tweets would be beneficial to stock analysis. The day the tweets were analyzed the stock of NXT-ID jumped barely over 1 percent, which explains the frequent mentions. Bloomberg was the news outlet that covered Trump's influence on the market, and they often make news articles regarding stock market movements.

Setup Sentiment Analyzers:

The sample set was created by hand judging randomly selecting tweets from 4/19 till 5/11. The sample set consists of 300 tweets labeled with a -1, 0, or 1. Negative tweets were denoted by '-1'; Neutral tweets were denoted by '0'; Positive tweets were denoted by '1'. The sample set does not come from an outside source because sets that judge sentiment often do not specifically analyze tweets of the stock market. This way, the sentiment analyzer could potentially understand more context behind each tweet in regards to the stock market.

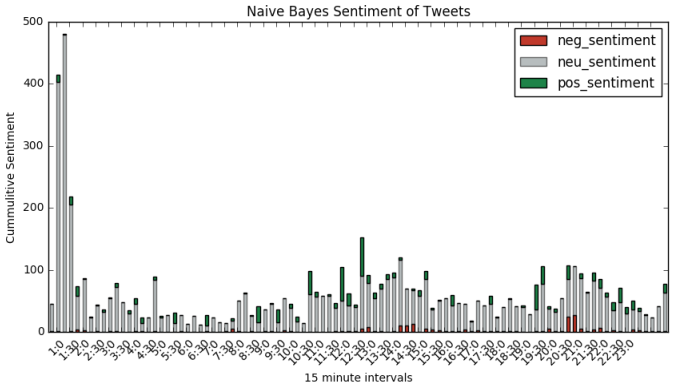
The Naive Bayes algorithm is a widely-used algorithm for text processing. Often, it can consider many parts of a text while also looking at the position of certain words as well. The Naive Bayes algorithm was used to train the sentiment analyzer, using 65% of the sample data set. The other 35% of the sample data set was used to test the algorithm. Here are the results:

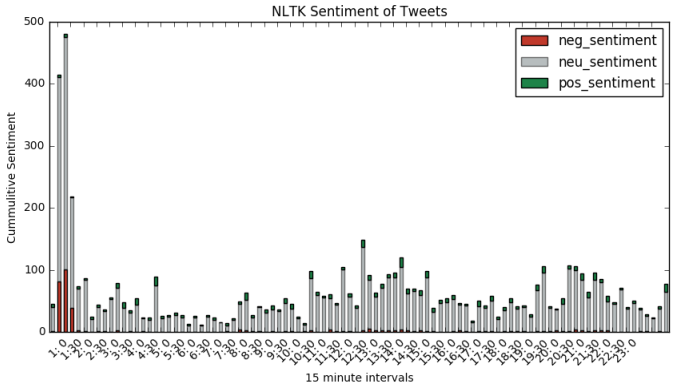
```
Accuracy: 0.6552901023890785
F-measure [-1]: 0.22727272727272727
F-measure [0]: 0.7673860911270983
F-measure [1]: 0.432
Precision [-1]: 0.4166666666666667
Precision [0]: 0.6866952789699571
Precision [1]: 0.5625
Recall [-1]: 0.15625
Recall [0]: 0.8695652173913043
Recall [1]: 0.35064935064935066
```

The overall accuracy is about 65%, which is average for a typical sentiment analyzer. The accuracy for negative tweets were very poor, however. This may be due to a lack of negative tweets within the sample. Most tweets are either neutral or positive, which explains why the accuracy is much better for the neutral and positive tweets.

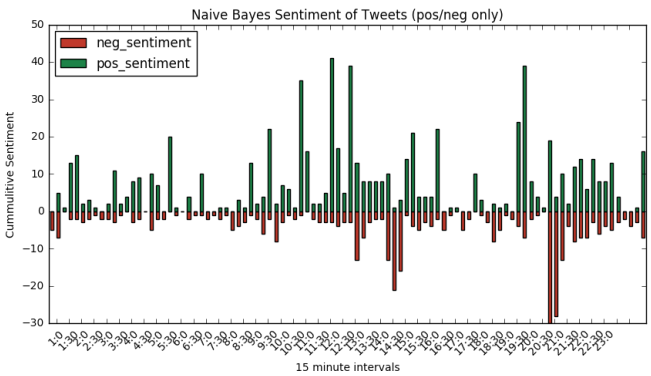
Display Sentiment & Relative stock price:

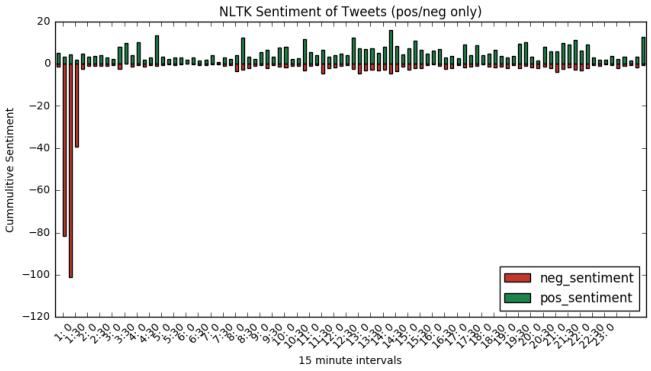
Below are graphs representing the cummulative sentiment of tweets in 15 minute intervals over the course of a day. Positive tweets are highlighted in red, grey neutral, and red negative. Oddly enough, the frequency of tweets spiked around 1:00 am. Often the spikes in frequency of tweets occur around 1:00 pm.

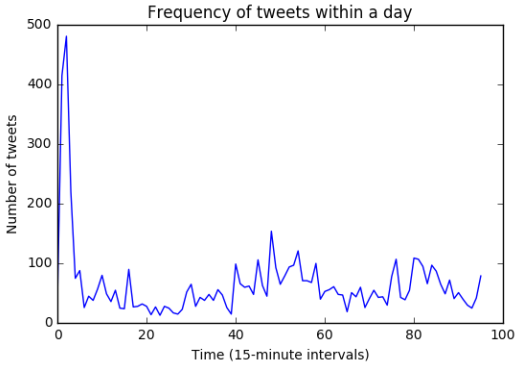




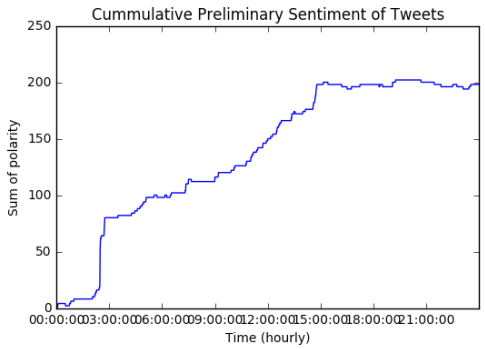
What's interesting is the market for the given day opened and immediately dropped. Potentially, the frequency or negativity of tweets at 1:00 am may have caused this. After the drop reached its lowest point, the stock market continued to rise until close. The sentiment analyzers seem to follow this trend by having majority positive tweets throughout the day.

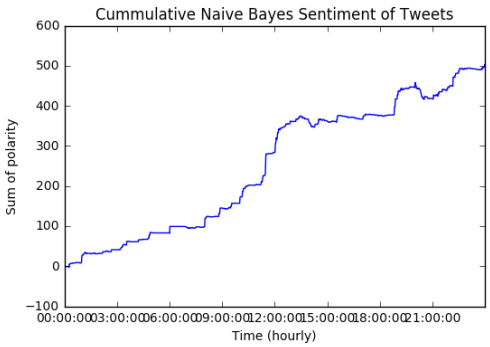


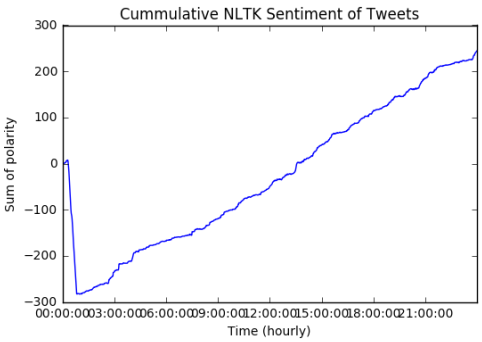




Generally, tweets are mostly positive or neutral, so the cumulative sentiment will often have a positive slope throughout the day. Typically, the stock market fluctuates a lot, so the cumulative sentiment of tweets does not identify minor fluctuations easily. However, Naive Bayes algorithm seems to do better than the others when identifying small chunks of negative sentiment throughout a day.







Early in the morning, two notable positive tweets were made. Both of whic h did not point to the stock price dropping heavily during market open. Mo re notable positive tweets were made around market close.

Notable tweets throughout the day:

```
236 retweets : @Un Rupok : 2017-05-11 20:39:08
    -RT @sphreco: How #blockchain tech will create a distribu
ted future for the energy sector: https://t.co/MvEDDVgD1Z via
@NASDAQ
```

```
200 retweets : @Nasdaq : 2017-05-11 14:34:05
    -RT @nvidia: Announcing NVIDIA Tesla #V100, the most advanced #AI GPU ever built.
Learn more: https://t.co/0mDV0LyFzd #GTC17 https://t.co/jw…

130 retweets : @Yuchenli : 2017-05-11 02:46:21
    -RT @RudyHavenstein: Nasdaq hit an all-time high!

Time for value investors to celebrate! https://t.co/kdMqM7c8wC

149 retweets : @Dale Vandenborre : 2017-05-11 00:32:08
    -RT @ahier: Stunning success from @Teladoc Q1 results
#telehealth #virtualcare $TDOC
via @Nasdaq

https://t.co/p7DuhadBPt
```

Conclusion:

The cumulative sentiment of tweets in 15 minute intervals seems to visually give more information than the cumulative sentiment of tweets throughout the entire day, probably because negative and positive tweets were easily visualized. In addition to that, the 15 minute intervals also give the number of tweets within 15 minute intervals, which can be extremely informative to how much the market may move. Generally, the cumulative sentiment of tweets has a tough time visualizing small fluctuations in the market, but different sentiment analyzers did well to predict large fluctuations within the market for the given day.

Tweets with more retweets did not carry a significant meaning in this experiment. Potentially, future trials may yield different results, but for now they do not relate to the stock market.

The negative sentiment of tweets accurately predicted the Nasdaq composite low of the day to be early in the market opening. Shortly, after the low sentiment analyzers saw overwhelmingly positive tweets, and the market price continued to rise till close.

If further analysis were to be done between tweets and the stock market, it would be beneficial to analyze a specific stock rather than a stock composite like Nasdaq. Many factors contribute to a stock price, and even more factors contribute to a collection of stock prices. Also, it may be beneficial to look at other forms of news and social media. Google search trends tend to lead many stock price changes. Studying that may yield interesting results as well.