Can Twitter provide enough information for predicting the stock market?

Maria Dolores Priego Porcuna

Introduction

Nowadays a huge percentage of financial companies are investing a lot of money on Social Media tools. They started focusing on this for some reasons. As we know, stock maket values depends on a lot of factors and trading algorithms are quite complex. Also, we cannot forget that information is power and that these financial companies based their investments on breaking news and social factors. They just started realizing about the important of applications like Twitter because of its huge free content. An maybe, over all of these data they can find a real good source, quick and powerful to retrieve information.

We are going to contrast this hypothesis. Does really Twitter contains information relevant for predicting the values of the stock markets? Also, for answering this question, we will need to find the most efficient way to retrieve and analyze this kind of information. Our study covers two important factors for the stock market, people's mood and breaking news. We will examine Twitter mood and mine news related to international markets. Our system separetes the logic for analyzing and retrieving data for these two areas, only at the end we will take a look at the big picture, and find some relevant patterns.

Our study is based on the recent success and media impact of this research area. We will contrast the hypothesis of recent research studies that have shown that Twitter's mood correlates somehow to DJIA closing values [1][2]. Moreover, we will try to perform some startups algorithms that are focused on Data Mining, performing algorithms that predicts newsworthy (to stock markets) tweets [3].

Data

For training and validation set we are going to focus on Aron Culotta's dataset (September 2009 - May 2010). Our aim is to build a functional and real-time system; therefore, we started collecting (properly) tweets from Twitter Streaming API on October 26th. For collecting these tweets with no cost we are using Amazon Web Services; EC2 where we have our scripts 24/7 hours running, and S3 where we back up our data. For periodic backup we use the third party tool s3sync.

We have registered our application at Twitter developers page in order to be able to be authorize to use the lastest streaming api version. At this point is easy to generate the request with the desired filters. As the sentimental and news systems are analyzed separately, when collecting the data we have designed different collectors (filters).

To begin with, for sentimental analysis we have been guided for Bollen et al. work [1] and we have filtered the data from Twitter's streaming in a similar way. We have selected tweets matching keywords: "i feel", "i am feeling", "i'm feeling", "i dont feel", "I'm", "Im", "I am", "makes me". So, the sentimental analysis is not going to be based on sentimental classification of a streaming of tweets, is much more simple. The system is based on a general classification of Twitter mood, not determining the sentiment of a tweet talking about things off-topic. It is straightforward parsing tweets of people that express how they feel. Here we assume that with those tweets we can analyze the relevant mood related to stock markets.

Second, for Data Mining we have performed different studies in order to find out which news affects Dow Jones closing values [4][5]. I already count on financial collaborators from CME group, who advice me to focus on news about The Federal Reserve or its president, Ben Bernanke. Also, they highlighted the importance of war news or government related news. This information was contrasted with some sources [6] and also with our dataset from 2009-2010, observing the effect of these news on Dow Jones values. Hence, we track tweets matching the following keywords "news fed", "news war", "news obama", "Bernanke news", "news goldman", "news government", and coming from important financial news resources like CNNMoney, Chicago Fed, NYFedResearch, DallasFed, New York Times, Wall Street Journal and The Economist.

Finally, for collecting the Dow Jones closing values automatically from the system, the third party tool, the gem 'yahoo_stock' is used.

Methods

A. Mood Analysis

This area is not new nor related to social media. A lot of trading processed have been based on people's sentiment during the history. Also, there are a lot of research studies that analyze the investor sentiment [7]. One important conclussion to highlight and take into account is the one made by Baker et al. [8] where they assert that global sentiment is a contrarian predictor of country-level returns. Therefore, I have chosen to focus only on a global market, Dow Jones Industrial Average and predict its volatility.

As Bollen et al. assert that Tense dimension has more correlation with this market than other POMS dimension [9], this study parse only the sentiment related to Calm dimension. Also we expect this to be a contrarion predictor referring to [8] work.

As POMS adjectives related to this are only a few unigrams, we have decided to take into account unigrams adjectives of fear from WordNet [10]. Our system score each tweet based on if it matches some of these adjectives. If it matches POMS adjectives as they demonstrate stronger sentiment feeling, it is going to be score with two points. On the other hand if it matches some of WordNet adjectives it is going to be score with one point. At the end of the day we will have a scoring grade based on how many times people expressed that they "are tense".

B. Mining News.

Our aim with this system is to be able to recognize when an important new comes out, so we can predict when the changes at DJIA is going to produce. Our hypothesis is that words within tweets in the past that were newsworthy have some resemblance to words within tweets that can be newsworthy now.

Our assumption is that SVM algorithm performs better than other Machine Learning algorithms at this task. Instead of implementing the algorithm from scratch, because of the timeline, I have decided to use the code from LIBSVM [12] which is an integrated software for support vector classification written by Chih-Chung Chang and Chih-Jen Lin, of National Taiwan University, Taipei.

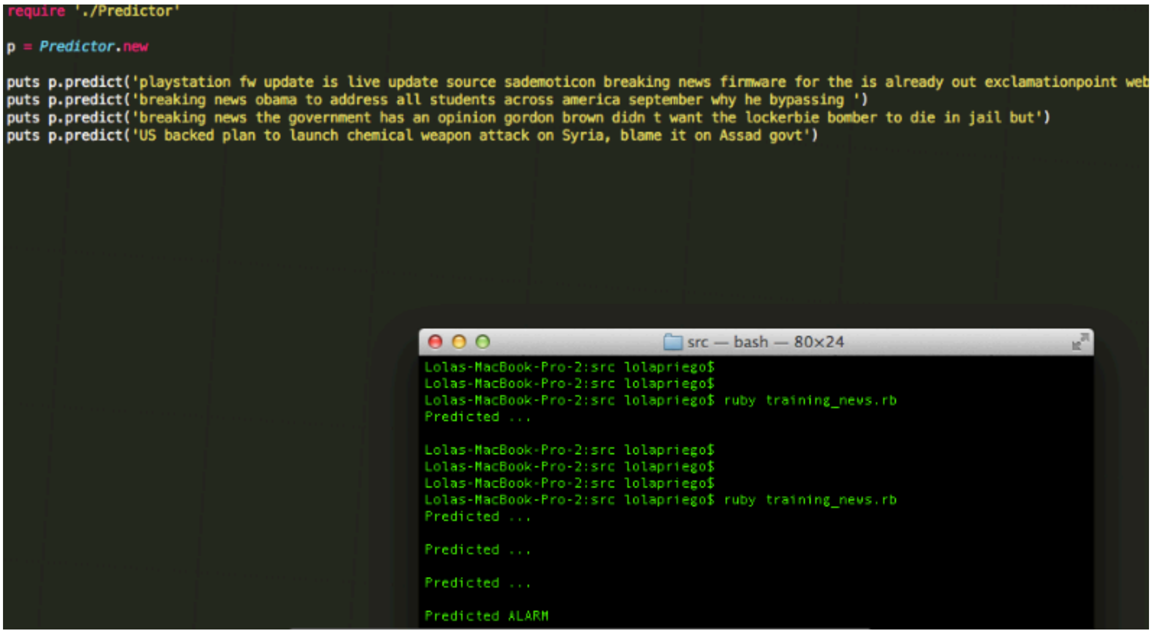
For training the algorithm, I looked at DJIA closing values within my training data period (September 2009 - May 2010) and anotating the days with an increment or decrement of 200 points, considering that this is a history significant movement in the market [13]. After that, I looked at the news that belongs to that specific days and labeled manually the news that seems to be newsworthy with the aid of a financial collaborator. From other days I hand-labeled the news that was completely irrelevant to the stock market or that was completely off-topic. Because of timeline, the algorithm of the system is not well-trained.

The criteria that I followed for selecting the algorithm parameters was selecting random parameters and test the algorithm with a validation set, and see which parameter performs better. Although exponential kernel, is more complex, performs worst than basic one. On the training set we obtained that basic kernel returns 60 alarms, #1, #2 and #3 returned 0 alarms and #4 90 alarms, within 60 days. For the error, the C parameter was selected as 2, but as soon as we do not put off range values, this does not look to affect that much. For the epsilon, a value of 0.0001 was chosen.

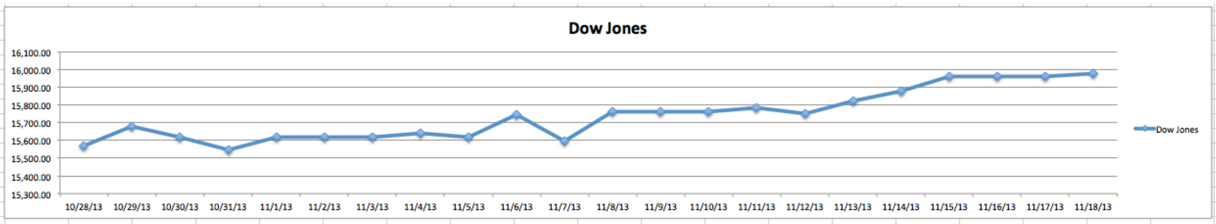
After having implemented the alarm system, is time to decide how this can affect to DJIA. In order to follow a similar logic to sentimental grading, we are going to score each alarm tweet. Each tweet categorized as relevant or alarm will be scored with one point. For example, if a success really important happens it impact on Twitter is going to be huge so a lot of tweets are going to talk about it. This means that if the system mistakenly tag a tweet as an alarm, but it is non relevant it only going to be scored with one point. On the other hand, if a tweet is really relevant (like Syria war) a lot of tweets will be categorized as alarms, and we will have more than ten points that day.

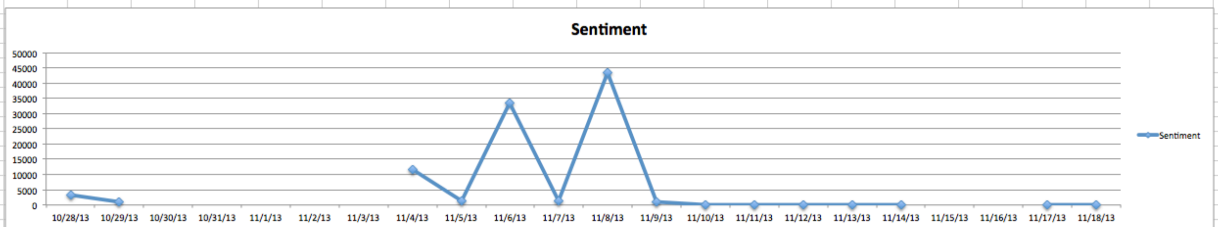
Experiments

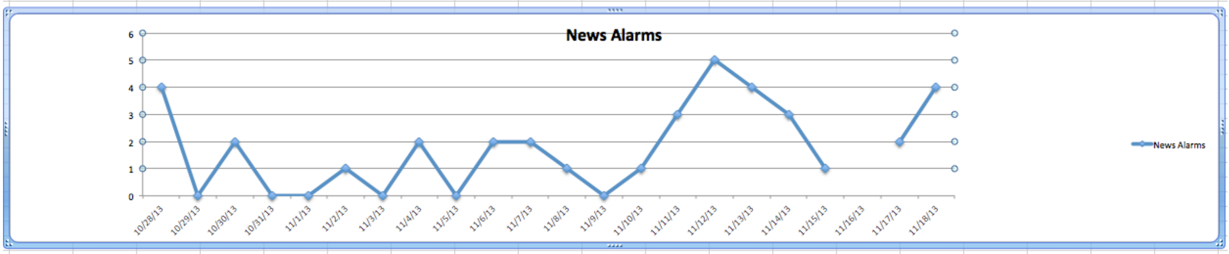
For testing the SVM algorithm, I took as a validation set news that belongs to historic relevant news, like 9/11 or news belonging current Twitter timeline, obtaining good results.



The final experiment consisted on testing the whole system. For the input data we grouped our dataset into separate files. Each one containing tweets within a day. We scored the sentiment tweets and the news tweets and obtained the next results







We can observe the blank space at the beginning of November at the sentiment classification. This was due to authentication issues with the Twitter API, and the amount of data collected. Also, the spike of November 12th and 13th of breaking news was due to false alarms like 'news war &' or 'Syria news war' or 'breaking news Obama just got a shower'. News containing characters on single words that did not appear into our SVM classifier makes it to crush. Therefore, it is quite important when monitoring this system to output the alarms produced in order to know if we are going in the right direction.
Another spike that we can observe is from 11/6 to 11/8 at calm dimension, that somehow correlates with the volatility of Dow Jones of that period.
Anyway, we should be careful when making conclusions about this experiment and its correlation with the Dow Jones closing values because of the lack of the volatility during this period (and its duration).

Related Work

Bollen et al. paper. This study is primarily based on this research paper, most of our assumptions are based on their hypothesis contrasted. We started from the point where after correlating DJIA with every POMS dimension, they observed that Calm dimension is more correlated. Also, they had a different scoring policy, because they increase the POMS adjectives by matching these ones with Google Queries dataset [14] (with a score related to how many times they have performed this query) which cost 150 usd, and we could not access.

Also, there is another study made by Zhang et al. [15] that performed the evaluation of hope, fear and worry over DJIA, S&P 500, and NASDAQ. They obtained good results, and inspired me to enlarge the study to the last two markets.

Finally, another system that guided my study was the startup Dataminr, that claimed that they detected a slight blip, linguistically on the stock market. Their algorithm *found that words within the tweet had some resemblance to tweets in the past that had turned out to be newsworthy, and that there was a clear immediate reaction to the tweet, though it had not yet rippled out to national news sources and market commentators*. Therefore, they are more focused on being the first resource, even previous to the news. Our algorithm is likely their algorithm, but our main source comes from news node in Twitter, so we are more focused on predicting a closing value rather than generating an immediate alarm.

Conclusions and Future Work

From the SVM algorithm test, we can deduce that, after improving and training it with a huge amount of news, the system will be quite capable of detecting newsworthy tweets. So, yes, it is possible to analyze efficiently the information contained in Twitter. In order to train efficiently the algorithm we could train it with all the news pre-filtered of the days with a 200 points increment. Also, it is important to train it with a huge amount of non-relevant news, which is one of the main reasons that cause false alarms.

From the sentiment experiments we can conclude, that it is more than a coincidence the spike produced at the beginning of November, which gives us green light to continue examining the correlation between DJIA and Twitter calmness.

The next step in the direction of experiments its normalize both scores, and sum it, with a proper grading policy, and perform the correlation between this score and DJIA, Nasdaq and S&P 500 closing values.

More features that can be included:

- Perform other Machine Learning algorithms, like Naive Bayes, to contrast the performance and see which algorithm performs better on a validation set. Also, I would like to perform Cross-Validation.
- Increase validation set for data miner system in order to select better parameters of SVM algorithm.
- If a new appears contaning just a single word, and this one does not belong to our training data at SVM, it will not be tagged as alarm.

[Check Github issues for more features and information, they are explained and tagged there].

References

1. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.
2. Bollen, J., Mao, H., & Pepe, A. (2011, July). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.
3. Startup Dataminr claims it gave investors a three-minute headstart to dump BlackBerry stock (http://gigaom.com/2013/11/04/startup-dataminr-claims-it-gave-investors-a-three-minute-headstart-to-dump-blackberry-stock/)
4. Dow Jones Industrial Average (http://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average)
5. Dow Jones Industrial Average, S&P 500 Post Biggest Losses of Year (http://abcnews.go.com/Business/bonds-stocks-affected-federal-reserves-announcement/story?id=19447197)
6. How Now, Dow? What Moves The DJIA? (http://www.investopedia.com/articles/02/082702.asp)
7. Brown, G. W., & Cliff, M. T. (2005). Investor Sentiment and Asset Valuation*.*The Journal of Business*, *78*(2), 405-440.
8. Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, local, and contagious investor sentiment. *Journal of Financial Economics*, *104*(2), 272-287.
9. Profile of Mood States (POMS). Jennifer A. Haythornthwaite, Ph.D. Robert R. Edwards, Ph.D. Department of Psychiatry & Behavioral Sciences. Johns Hopkins University. Baltimore, Maryland USA
10. WordNet-Affect (http://wndomains.fbk.eu/wnaffect.html)
11. Support Vector Machine (http://en.wikipedia.org/wiki/Support_vector_machine)
12. LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)
13. ECONOMIC CONSEQUENCES of WAR on the U.S. ECONOMY(http://www.thereformedbroker.com/wp-content/uploads/2012/02/Economic-Consequences-of-War.pdf)
14. Google Queries Scores (http://catalog.ldc.upenn.edu/LDC2006T13)
15. Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear" Procedia-Social and Behavioral Sciences, 26, 55-62. (http://www.sciencedirect.com/science/article/pii/S1877042811023895)
16. Dataminr (http://www.dataminr.com/)