Analysing news articles impact in stock price movements

Marcelo Grossi

August 8, 2016

Abstract

Stock market prediction has been a topic of great interest for researchers. Numerous attempts to "beat the market" have been made without been able to consistently and accurately predict the movement of stock prices.

In recent years, following the increase in computational processing capabilities, researchers have been studying the relationship between news articles and stock price movement (Fu et al., 2008; Schumaker and Chen, 2009). Most methods are based on sentiment analysis, where the news content is classified according to its potential of moving the underlying stock's prices up (good news) or down (bad news). These methods however, do not take into account the financial instrument's context in determining the content of the news article, which greatly impairs the results of such studies.

This work assumes the efficient market hypothesis which dictates that the stock's prices reflects all the information available (including historical prices, public news information and even insider information) (Fama, 1965) and that everyone has some degree of access to the information. The relationship between closing historical prices from several companies will be related to relevant information in news articles by identifying the most important terms in a daily period and testing for some sort of causality between the significant terms and the price time series. This technique allows for direct analysis between aggregated news information and stock data, greatly reducing the complexity of dealing with textual information.

1 Introduction

The relationship between news articles and stock price movements has been studied for several years (Gidfalvi, 2001; Fu et al., 2008). The most prevalent methodology of relating news articles to stock prices is by capturing the investor sentiment (Mitra and Mitra, 2011) from the textual information (if it is positive, negative or neutral) and assigning a probability of whether the stock prices were likely to move upwards, downwards or stay the same. Previous researches (Tetlock, 2007) have shown that negative sentiment predicts downward pressure

on market prices, and also (Barber and Odean, 2008) that significant news will affect the traders beliefs, what translates into an increase in trading volume.

Sentiment can also be understood as the informational value of news and can be interpreted in several ways, such as via the use of financial dictionaries for analysing positive terms versus negative terms i.e., 'profit' and 'exceeds' versus 'bankrupt' and 'loss'; or via machine learning and natural language processing techniques to automatically generate lexicons for good and bad news sentiment (Oliveira et al., 2014), achieving excellent results compared to baseline lexicons.

By using a simple naïve classifier, (Gidfalvi, 2001) finds definite predictive power for stock price movement within a 20 minute window before and after a news article becomes publicly available. These sentiment-based researches are very promising, and although show high correlation with stock indicator movement it is unclear how this relationship comes to be.

This paper tries to contribute to the work by introducing a different approach to the analysis of the relationship between relevant terms in a textual corpus and stock prices. By computing the relative importance of a term over time using the Term Frequency Inverse Document Frequency (Ramos, 2003) weight, it allows for the ability to use regular tools and models for time series analysis which abstracts away the difficulties of dealing with textual data directly, as textual information is now effectively converted pragmatically into a usable time series.

To evaluate this model a Granger causality test (Granger, 1969, 1980) was used to assess the predictive power of the news time series over the stock price series. This test is performed between the top n most relevant terms of the day against the full set of stock prices. So the question posed is inverted; instead of asking 'does this piece of news influence my stock?' we assume that news has influence over prices, and try to find 'what stocks will be influenced by the current news state?'.

After selecting the term t, and stock s tuples for a given day d a naïve model is tested whereby the price of day $d + 1$ is predicted and compared with the actual price.

2 Data sources

Online news articles from the New York Times (`http://www.nytimes.com/`) were used as the textual corpus of this research. Around 50,000 business-related news articles were scraped from the period between January 1st, 2013 and October 19th, 2014. As one of the most respected news media outlets in the world, the New York Times should provide a sufficiently good breadth of market relevant information for analysing news impact on stock prices. More specialized news providers (such as Bloomberg or Financial Times) that could potentially outperform the current data source were not considered due to their lack of impact on small investors.

In order to maximize the expectation of news having an influence on stock prices, this data had to come from the same region as the news provider. Therefore, only stocks traded in the New York Stock Exchange were considered and

daily price from twenty nine blue chip companies was gathered for a period that encompasses completely the news articles' database date range.

3 From text to time series

Each news article a was converted from raw unstructured textual data to a computer-friendly representation, in the form of a vector of term and related frequency f, where a term t is a word n-gram (for this experiment *1* and *2*-grams were used). This format is also commonly known as *bag-of-words*.

$$a = [(t_1, f_1), (t_2, f_2), .., (t_n, f_n)]$$

To transform the unstructured text into the *bag-of-words* format a series of sequential transformations was applied. Contractions are expanded, i.e, *it's* $\rightarrow$ *it is*; most common words in the English language are removed (also known as *stop words*); sentences are separated (so $n-$grams bigger than unity can be generated only if they came from the same sentence); word tokenization (transform sentence s in a collection of words W_s) is used; part-of-speech tagging (or $POST$) was performed in each sentence (disambiguate each word's grammatical function, i.e., adjective, subject, verb, etc.); word lemmatization (use $POST$ extracted at the previous step to calculate the word's lemma, i.e., *saw, verb* $\rightarrow$ *see* or *saw, subject* $\rightarrow$ *saw*); and finally generation of $n-$grams – i.e., given that $W_{s_1} = [w_1, w_2, .., w_n]$ is a set of words from sentence one, the *2*-grams collection would be of the form $[w_1w_2, w_2w_3, .., w_{n-1}w_n]$.

Each news article (converted to *bag-of-words*) $a \in A$ was grouped by the date d it was published, and a daily summary was produced whereby a score was assigned to each $n-$gram following the Term Frequency Inverse Document Frequency algorithm. This score is calculated by multiplying the $n-$gram frequency (number of times $n-$gram t appears in all of A_d over the total number of $n-$grams in A_d) by the logarithm of the inverse of the document frequency (total number of articles in the day $|A_d|$ over the number of articles $|A_t|$ that $n-$gram t appears in). Another way of understanding the $tfidf_{t_d}$ is that it expresses the relative importance of $n-$gram t over the course of day d. So even if a term (or $n-$gram) appears in most documents with very high frequency (high tf), it will get heavily penalized by the idf, where $0 \leq tfidf_t$. Although the score is unbounded in its upper limits, it is expected that in an unbiased big collection of articles, this value stays close to zero (this is due to the normalization of the tf whereby the individual frequencies are divided by the total frequency of the day). From the daily summaries obtained through $tfidf$ it becomes straight forward to calculate a time series TS for any $n-$gram t given a set of all days $D_{start \rightarrow end}$ between days d_{start} and d_{end}.

$$TS_t = \forall tfidf_{t_d},$$

where $d \in D_{start \rightarrow end}$.

4 Selecting terms and stocks

One of the objectives of this research is to find some relationship between the daily stock prices and the time series produced from the textual corpus. A brute force attempt will not be attempted, as there are roughly 30,000 different $n-$grams each day and those in the lower end of the $TFIDF$ weighing scheme are either too frequent to be of any use, or too infrequent so as to not have enough data to be relevant. For this study, an arbitrary amount of 100 $n-$grams with the best score are selected from each day's summary, and a time series is extracted for each one. This set will henceforth be mentioned as T_d.

The set of stock price data S_d, as a dependent variable, is a slice from the complete stock price information from January 1$^{\text{st}}$, 2013 to d; where d is the day being evaluated for the independent variables in T_d. Interpolation is performed in S_d at this point, to make sure all news data points are captured and potential relevant information published on weekends or holidays is properly modelled.

In order to produce the final term $t_d \in T_d$, stock $s_d \in S_d$ tuples a test must be performed to validate wether or not t_d has any prediction power over s_d. With that same objective, (Granger, 1969, 1980) proposed a *predictive causality* statistical test to determine if a time series is useful in forecasting another. This *Granger-causality test*, or simply *Granger test* is based on two fundamental principles according to (Eichler, 2012):

1. "the effect does not precede its cause in time";
2. "the cause has unique information about the series being caused that is not available otherwise";

Following from these principles, it can then be formulated by letting y and x be two stationary time series, where the null hypothesis is that x does not *Granger-cause* y, if the *F-test* (or *Wald test*) between a univariate auto-regression of y (1) and a multivariate auto-regression of both y and x (2) fails to find any explanatory power after the exogenous variable x is added to the model.

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + .. + \beta_n y_{t-n} + \varepsilon_t \tag{1}$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + .. + \beta_n y_{t-n} + \alpha_1 x_{t-1} + .. + \alpha_n x_{t-n} + \varepsilon_t \tag{2}$$

It is not uncommon to find a bi-directional causality relationship between the variables when *Granger-testing*, therefore both directions are tested and if no causality is found, or the causality is found in both directions the tuple is disregarded. Furhter information on the actual methodology used for testing *Granger-causality* in this study can be found in (Toda and Yamamoto, 1995), or excellently summarized in (Giles, 2011).

The resulting tuples (t_d, s_d) where $t_d \xrightarrow[causes]{Granger} s_d$ (and $t_d \not\xrightarrow[causes]{Granger} s_d$) can now be used to test the effectiveness of the predicting power of t_d over s_d. As an example 1 shows a visualization of terms t_d related to stocks s_d for

$d = 14/08/2014$ (this day was chosen to allow a $9:1$ split of the dataset, saving 10% for the tests).

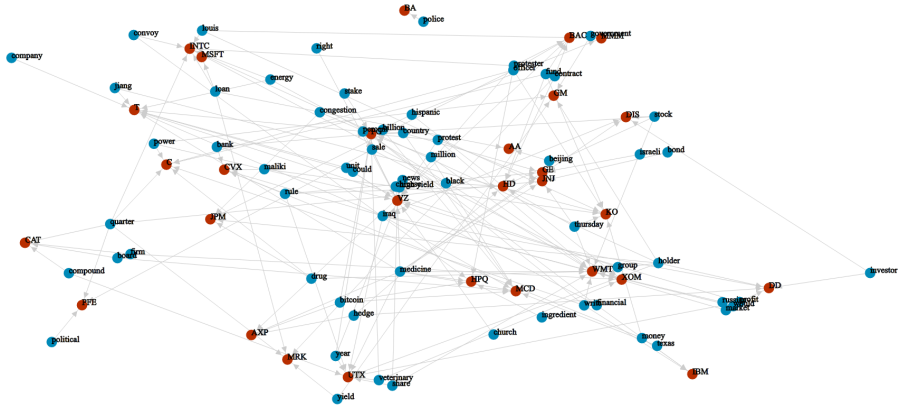

Figure 1: Directed graph of *Granger-causing* terms (*blue*) and stock symbols (*brown*) for the sample period 01/01/2013 to 14/08/2014

5 Predicting price movements

To validate the generated term predicting time series an *autoregressive integrated moving average* (*ARIMA*) model is fit to each term, stock and day tuples, and is first individually compared against the actual stock prices (where a sum of squared errors is computed). Afterwards all the predictions for the same stock on the same day are aggregated and an average prediction is generated. This aggregated prediction's sum of squared errors are then compared to the best sum of square errors of the previous step in order to assess if a composite prediction is better than the best individual one. This test is repeated for all days of the test dataset (namely, the interval between 14/08/2014 to 19/09/2014).

Do same ARIMA fitting but for all precitors together?

A tradeability test is also performed to verify if this technique can actually be used as a valid investment strategy. This test is a simple directional correctness measure. If for day d the predictions (best individual and aggregated) are able to correctly forecast the direction in which the prices have moved the next day $d+1$, it is considered correct. If the correctness rate is found to be better than random (far from 50%), it is considered a tradeable strategy. Note that even if the correctness rate is close 0% it is considered tradeable, as all an investor has to do is invert the signal.

6 Results

The tuples generated for *Granger-causing* terms and stocks were quite interesting. The term *'loan'* was matched with stock prices from *Alcoa*, *Intel*, *Merck & Co.*, and *Microsoft* indicating a predictive relationship between the historic relative importance of this term in the New York Times media coverage news corpus and the aforementioned companies' stock prices, as traded in the New York Stock Exchange.

The philosophy of why this term is related to these particular stocks is outside the scope of this research but other more peculiar terms were also matched with several stocks. Terms such as 'billion', 'million', 'could', 'year' and 'thursday' were inexplicably found to have a *Granger-causal* relationship with many blue chip stock prices. Given the common and widespread usage of these terms, they may have biased the *TFIDF* score making the interpretation of their relative importance into a proxy of some other generating factor - perhaps the volume of news produced? Or something else entirely.

- results for the validation section here - - results with sum of square errors (compare normal arima versus arima with exogenous) - show probability of directional correctness (predicted to go up/down versus actually went up/down) - show average predictions of multiple term time series related to same stock and take average as a consolidated prediction. - show results with square error - show probability of directional correctness

7 Conclusion

A different approach of relating textual data and stock price movements is introduced and ... **- small introduction of what was done, show results and come up with an explanation of why they were good/bad.**

Some caveats are important to be mentioned. Using a larger dataset, that spanned multiple years and sourced from different providers, would help offset potential editorial biases and increase variance. Also, if both the time and date of publishing were available for each news article, an equivalent intra-day research could show different insights into news and stock relationship.

Future research may include repeating this same experiments using different textual filtering options, such as capturing only nouns or named entities from the textual corpus to avoid some of the more peculiar results founds here. Another avenue of study can be to compare the effectiveness of different news sources; or build a directed term to stock graph and procure unknown relationships between stocks on different sectors or with weak intuitive relationships. Experimenting with different time series models and comparing results can also be a very interesting path to pursue.

References

T. Fu, K. Lee, D. Sze, F. Chung, and C. Ng, "Discovering the correlation between stock time series and financial news," in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Institute of Electrical & Electronics Engineers (IEEE), dec 2008.

R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news," *ACM Transactions on Information Systems*, vol. 27, no. 2, pp. 1–19, feb 2009.

E. F. Fama, "The behavior of stock-market prices," *Journal of business*, pp. 34–105, 1965.

G. Gidfalvi, "Using news articles to predict stock price movements," Department of Computer Science and Engineering, University of California, Tech. Rep., 2001.

G. Mitra and L. Mitra, Eds., *The Handbook of News Analytics in Finance*. Wiley, 2011.

P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

B. M. Barber and T. Odean, "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors," *Review of Financial Studies*, vol. 21, no. 2, pp. 785–818, 2008.

N. Oliveira, P. Cortez, and N. Areal, "Automatic creation of stock market lexicons for sentiment analysis using StockTwits data," in *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS'14*. Association for Computing Machinery (ACM), 2014.

J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003.

C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, p. 424, aug 1969.

——, "Testing for causality," *Journal of Economic Dynamics and Control*, vol. 2, pp. 329–352, jan 1980.

M. Eichler, "Causal inference in time series analysis," *Causality: Statistical perspectives and applications*, pp. 327–354, 2012.

H. Y. Toda and T. Yamamoto, "Statistical inference in vector autoregressions with possibly integrated processes," *Journal of Econometrics*, vol. 66, no. 1-2, pp. 225–250, mar 1995.

D. Giles, "Testing for granger causality," http://davegiles.blogspot.co.uk/2011/04/testing-for-granger-causality.html, apr 2011.