Analysing news articles impact in stock price movements

Marcelo Grossi

August 2, 2016

Abstract

Stock market prediction has always 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 stocks 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 strong 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. We will look at historical prices from different sources/indices and investigate the relationship between stock prices and news articles by identifying the important events in them and trying to match the news from that period with important price movements. This technique will allow us to identify for instance, which terminology or characteristic of news item has good or bad connotations to the price and how much impact can be attributed to it.

As recent terrorist events have shown, some tend to be local with regard to stock market impact, others have more global ramifications. This technique allow us to differentiate and localize the impact of a given new piece. Many other applications for this work can be thought of including, but not limited to a trade system that tries to anticipate price movements based on news repercussion and grouping of stocks that have similar news influencers but are not necessarily in the same market sector (discovery of non-obvious stock clustering).

1 Introduction

The relationship between news articles and stock price movements has been studied for several years (?). 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 (?) have shown more downside impact then upside in relation to the attributed news sentiment (i.e., bad news has more negative impact than good news has positive impact).

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 (Oliveira et al., 2014); or a mixed model (?) whereby the researchers build a sentiment index that can be correlated to an underlying asset (?).

¡¡¡¡¡¡¡¡¡¡¡¡¡¡ ¡¡¡ WHICH ONES GIVE WHICH RESULTS AND HOW SUCCESSFUL AT PREDICTION ? ¿¿¿ These sentiment-based techniques are very promising. By using a simple naïive 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.

and show strong correlation with stock price movement, but it is unclear how this relationship comes to be. ¿¿¿¿¿¿¿¿¿¿¿¿¿¿¿

This paper tries to contribute to the work with 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 (?) weights, it allows for the ability to use regular tools for time series analysis which abstracts away the difficulties of dealing with textual data directly.

To evaluate this model a Granger causality test (?) 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 on the New York Stock Exchange were considered and daily price information 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 (*TF-IDF*) 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 $D_{start \to end}$ between days d_{start} and d_{end}.

$$TS_t = \forall tfidf_{t_{d_i}},$$

where $d_i \in D_{start \to end}$.

4 Evaluation

tell how granger test works.

tell how to match term ts and stock.

present results with examples of stocks and terms that were found.

do prediction of day + 1 using the news relationship using at least two methods (arima and pattern dtw).

compare predictions with baseline index.

5 Conclusion

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

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.

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