Relating news articles summaries to stock prices

Marcelo Grossi
School of Computing
Dublin City University
Dublin, Ireland
Email: marcelo.grossi2@mail.dcu.ie

***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 [1, 2]. 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. And also it is cumbersome to work with textual data in relation to numerical time series data.

This work proposes a novel approach to relating textual data to stock price movements by identifying the most important terms in the daily corpus of news articles and testing for a Granger-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.

To conclude, this study further analyses this relationship between news articles and stock price movements by using the first as regressors for predicting the latter and finds that there is a clear advantage of using all stock-related terms as covariates for predicting the target stocks price movements. Despite of the better predictions made by using the textual data, it is also found that these forecasts are not viable as a direct trading strategy.

I. Introduction

The relationship between news articles and stock price movements has been studied for several years [3, 1]. The most prevalent methodology of relating news articles to stock prices is by capturing the investor sentiment [4] 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 [5] have shown that negative sentiment predicts downward pressure on market prices, and also [6] 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 [7], achieving excellent results compared to baseline lexicons.

By using a simple naïve classifier, Gidfalvi [3] 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 analyses are very promising, and although they show high correlation with stock indicator movement it is difficult to work with textual data directly, regardless of the approach used. A general classification of sentiment (be that through dictionary/lexicon techniques or machine learning) does not take into account each stock's individual context. An extreme example would be the term 'bankruptcy' which might be classified as negative in every scenario, but if a company's business is 'bankruptcy attorneys' this term being on the rise might actually mean something good, as it can relate to more potential business available.

This paper contributes to the area with a novel 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 [8]

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 [9, 10] 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 an 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 ARIMAX model is tested whereby the price of day $d+1$ is predicted and directional correctness and RMSE key performance indicators are calculated and compared.

II. 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.

A time series is a sequence of numerical data points in time successive order recorded at regular intervals. The gathered stock prices information is comprised of one daily time series of closing stock prices for each blue chip company.

III. From text to time series

One difficulty of working with time series and textual data is that the latter does not conform to the time series format. The data is not numerical and it is usually interpreted as one isolated event in time (usually time of publication) with no direct relationship to other textual pieces. This work brings a novel approach of converting textual data into usable time series, allowing regular time series analyses models to be applied directly.

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 were 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 in 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 (*tfidf*).

$$TS_t = (tfidf_{t_d}, \forall d \in D)$$

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.

IV. Selecting terms and stocks

One of the objectives of this research is to find if there is a relationship between the daily stock prices and the time series produced from the textual corpus. A brute force approach 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 known 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 1st, 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 [9, 10] 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 [11]:

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, the *Granger test* model 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 (equation 1) and a multivariate auto-regression of both y and x (equation 2) fails to find any explanatory power after the exogenous variable x is added to the model.

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

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

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 [12], or excellently summarized in [13].

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. A sample of such relationship can be seen in figure 1 where terms related to stocks 'KO' (*Coca-Cola Co.*) and 'PFE' (*Pfizer Inc.*) are shown for $d = 14/08/2014$ (this day was chosen to allow a $9:1$ split of the dataset, saving 10% for the tests).

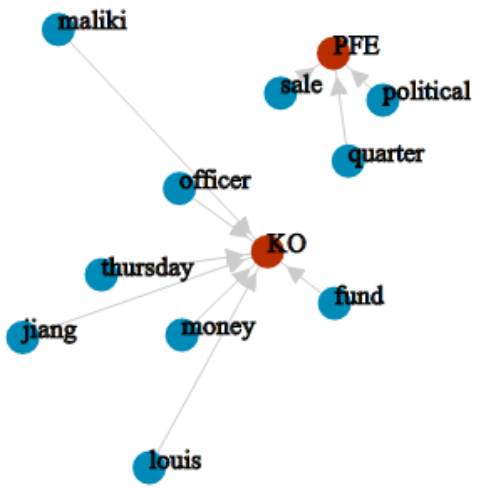

Fig. 1. Directed graph sample of *Granger-causing* terms (*blue*) and stock symbols (*brown*)

V. Predicting price movements

To validate the predicting potential of the generated term time series, a model is chosen to fit the data and forecast out-of-sample prices for the stocks using the term time series as regressors. These results are then compared to the same model without any regressors to assess whether or not the term time series has an influence on stock prices (if prediction made with terms is better than without the terms it becomes patent it's influence). To finalize, a test to measure the ability of

the predictions to be used as an investment strategy is proposed in order to evaluate the usefulness of this model in real world scenarios.

Widely used in time series analysis, and first introduced by Box and Jenkins in 1976 (recently reviewed in [14]) the *autoregressive integrated moving average with exogenous variable* (*ARIMAX*) model is fit to each term, stock tuples, where the stock price series is the target function and the term series functions as a predictor (or exogenous variable). The tests are performed using each term series as an exogenous variable individually, then all term series related to a stock are used as exogenous variables to fit the same model. A final model is fit to the stock prices without any predictors, to serve as the baseline for further comparisons. These models are henceforth referred to as the individual (Ind), aggregated (Agg) and no-exogenous models ($NoEx$).

The *ARIMAX* model, which is an integrated variation of the *ARMA* model with exogenous variables, is fit for modelling auto-correlated data with or without seasonality components (in this study only the non-seasonal variety is used), where the integrated part is a shorthand for differencing the series before calculating the coefficients to ensure stationary. The introduction of exogenous variables in the model involves the addition of the covariate x_t to the right hand side of the *ARMA* formula. Equations 3 and 4 show the *ARMA(p,q)* and *ARMAX(p,q)* models respectively, where z_t is a white noise process.

$$y_t = \phi_1 y_{t-1} + .. + \phi_p y_{t-p} - \theta_1 z_{t-1} - .. - \theta_q z_{t-q} + z_t \quad (3)$$

$$y_t = \beta x_t + \phi_1 y_{t-1} + .. + \phi_p y_{t-p} - \theta_1 z_{t-1} - .. - \theta_q z_{t-q} + z_t \quad (4)$$

To asses the quality of time series forecasts, it is usual to calculate the root mean squared errors (or *RMSE*) and compare it against other RMSE's produced by different models on the same dataset, the model with lowest RMSE performs best. As the objective in this study is to compare the forecast quality across different time series (with potentially different price intervals), a scaling function is applied to the stock price series before calculating the RMSE (a scaled series can be understood in this context as a time series whose values lie within the $[0, 1]$ interval, and equation 5 shows the scaling function $SCALE_s$ for stock series s).

$$SCALE_s = \frac{s - \min s}{\max\{s - \min s\}} \quad (5)$$

Lastly, a tradeability test is performed to verify if this technique can be used directly as a viable investment strategy. Several approaches can be used to assess tradeability, the most usual being the construction of a full trading strategy surrounding the signals produced by the model. This can be very cumbersome and prone to masking the real performance of the underlying signal generating model, by use of clever investment strategies one can produce profits that are completely unrelated to the model being tested.

To avoid this pitfall, a simple directional correctness ratio measure is introduced where if for day d the predictions (individual, aggregated and no-exogenous) are able to correctly forecast the direction in which the prices have moved the following 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 viable strategy. Note that even if the correctness rate is far below 50% it can still be considered viable, as all an investor has to do is invert the signal.

These tests are computed for all days of the test dataset (namely, the interval between 14/08/2014 and 20/09/2014).

VI. Results

¡TODO:MORE CONCLUSIONS CAN BE DRAWN FROM THIS?¿ 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.

The individual and no-exogenous models performed similarly, with a slight edge to the individual models. The aggregated models performed best in both measures ($RMSE$ and $directionalcorrectness$). Although there

is a definitive improvement from the no-exogenous models to their exogenous counterparts, no model has showed significant tradeability potential.

The *ARIMAX* model turned out to be a poor predictor of the stock series, failing to fit the data exceptionally well. It would be interesting to analyse the same dataset with a collection of more advanced time series models. Such exploration work will not be performed here as this model was merely an example of what can be done with the generated term time series, and also to compare its performance against that of the no-exogenous model.

The directional correctness ratio indicator showed a variation between 35% and 63% correctness, with an average of 49.9% and variance of 0.43%. This result, coupled with a t-test (with 5% significance) indicates that the predictions made by the model are no better than random and that using this model with the term time series is not directly translatable into winning trading strategies. Figure 2 plots the directional correctness binned frequencies and it shows quite clearly the aggregate models' superiority if compared to the individual and no-exogenous models.

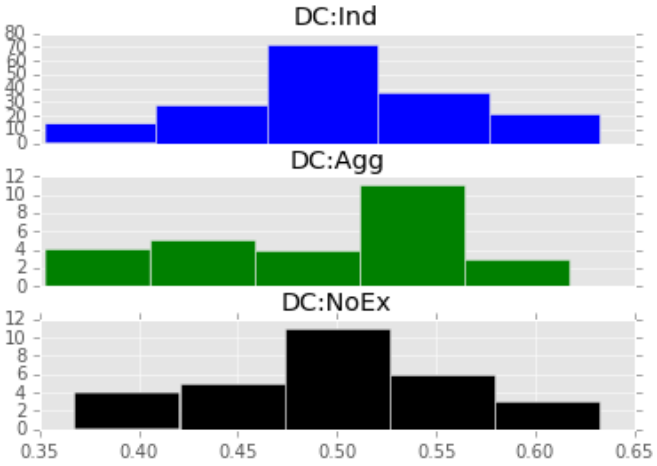

Fig. 2. Directional correctness ratio of forecasts binned frequencies

The best directional correctness results were for the stock symbols *KO* (Coca-Cola Co., using exogenous series from term *"money"*) and *PFE* (Pfizer Inc., using exogenous series from term *"quarter"*) with 35% and 63% correctness respectively. Inverting the directional predictions of Coca-Cola Co. would yield a correctness rate of 65%, becoming slightly better than Pfizer's.

The root mean square error measure allows for a direct comparison between the individual, aggregated and no-exogenous models. The results in this area corroborate the directional correctness results, showing that in general, the aggregated model fairs better than the individual and no-exogenous models. This clearly indicates that using all available information is more advantageous than modelling the terms individually and also than not using the term time series at all. The descriptive statistics of this indicator can seen in table I and figure 3.

TABLE I
ROOT MEAN SQUARE ERROR

	Obs	Min	Max	μ	σ^2
All	219	0.035	0.584	0.148	0.007
Individual	165	0.043	0.583	0.148	0.007
Aggregated	26	0.056	0.322	0.141	0.006
No-exogenous	28	0.035	0.584	0.153	0.012

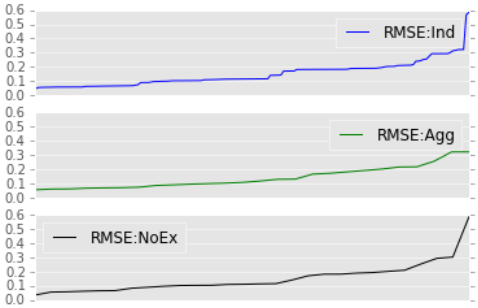

Fig. 3. Root mean square error for individual, aggregated and no-exogenous models

For more detailed information, the complete research, code and results can be found at https://github.com/magrossi/news_n_stocks.

VII. Conclusion

Using public and free data sources, a different approach of relating textual data and stock price movements is introduced by generating term time series data pragmatically from $TFIDF$ scores. This novel technique allows for a fully contextualized analysis of textual and stock data, and eliminates the cumbersomeness of dealing with textual data directly.

A validation model for the ability of the generated term time series to forecast stock price movements was tested and although it did not show direct tradeability capabilities, it clearly indicated that modelling stock prices using all related term time series is more advantageous and should encourage investors to use textual information in their predicting models to increase the quality of their output trading signals.

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 of different sectors or with weak intuitive relationships. Experimenting with different time series models such as transfer models, VAR, VECM, and others, then compare results can also be a very interesting research path.

REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

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

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

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

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

[12] 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.

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

[14] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis : forecasting and control*. Hoboken, N.J: John Wiley, 2008.