Summary about the paper

Robert P. Schumaker and Hsinchun Chen.

TEXTUAL ANALYSIS OF STOCK MARKET PREDICTION USING BREAKING FINANTIAL NEWS:

The AZFinText System

Maths for Big Data

GAEL RUTA GATERA & EDUARD CALSINA PLA & GENÍS LÁINEZ MORENO,

Univertitat Autònoma de Barcelona. 18/05/2020

Abstract

This report is written as a review to the "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System" article published in ACM Transactions on Information Systems (Schumaker and Chen, 2009). The prediction of price movement of Stocks(securities) has been an interesting field of study since the inception of exchanges such as the Chicago Mercantile Exchange (CME), New York Stock Exchange (NYSE), Nasdaq and others for many reasons. From a financial perspective, due to large sums of money companies and hedge funds can either make or loose in single trades especially with margin trading. From a mathematical research stand point, where there is a challenge in modelling the stochastic and often volatile nature assets. In this case, from a financial text mining perspective as quarterly reports or breaking news stories can dramatically affect the share price of a security. As stated in their abstract they tried to create a "predictive machine learning approach for financial news articles analysis using several different textual representations: Bag of Words, Noun Phrases, and Named Entities. (Schumaker and Chen, 2009)"

1 Introduction

The majority of literature on financial text mining relies on identifying a predefined set of keywords and machine learning techniques. These methods typically assign weights to keywords in proportion to the movement of a share price. A great example of when news had a negative impact on the price of an asset is when Ethiopian Airline's Boeing 787 (Flight ET 302) crashed on March 10th 2019 (Korte, 2019). Hours after the crash, when details emerged and many news articles accusing Boeing of faulty engineering and maintenance, the stock value of Boeing decreased significantly in subsequent days. The following week CNN released an article stating "Shares of Boeing were down another 1.4% on Friday and were eyeing weekly losses of around 12%. That's equivalent to roughly $ 28 billion in lost market share for the aerospace and defense contractor.(Bourgi, 2020)"

In this research's textual representation, they investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five week period. They applied the analysis to estimate a discrete stock price twenty minutes after a news article was released. Using a Support Vector Machine (SVM) derivative specially tailored for discrete numeric prediction and models containing different stock-specific variables,

2 Literature Review

The theories that try to predict the future prices of Stock Market securities are Efficient Market Hypothesis (EMH) and Random walk theory (Fama, 2016). In EMH, it is assumed that the price of a security reflects all of the information available and that everyone has some degree of access to the information. Fama's theory further breaks EMH into three forms: Weak, Semi-Strong, and Strong (Schumaker and Chen, 2009). On the other hand Random Walk Theory has similar theoretical underpinnings to Semi-Strong EMH where all public information is assumed to be available to everyone. However, this theory declares that even with such information, future prediction is ineffective.

The two aforementioned theories gave rise to two elemental and prominent trading philosophies; Fundamental and Technical Analysis (Majaski, 2020). Fundamental analysis concerns itself with analyzing the financial numbers derived from sources such as the overall economy, a particular industry's sector, or most typically, from the company itself. In Forex exchanges for example indicators such as inflation, joblessness, Interest Rates and national debt levels can all play a part in determining the price of a traded currency (Tajik, 2019). If the indicators are performing poorly then price of the currency will go down and vice-versa. Technical analysis on the other hand, evaluates the prices based on historical and time-series data represented by candles and charts patterns. Common methodology is that both Technical and Fundamental analysts share is using financial news articles such as the one we previously mentioned to predict price movements due to their significance.

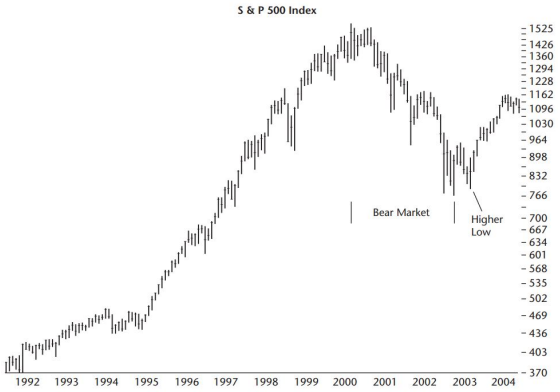

Figure 1: Standard & Poor's 500 stock index from 1991 to 2004

This research mentioned an external project conducted by LeBaron et. al. that used a simulated stock market with simulated traders that followed certain rules responding to changes in the market with each trader varyied on the timing between the point of receiving the information and reacting to it (Lebaron et al., 1999). The results of the simulations indicated that simulated traders that acted quickly formed technical analysis while traders that possessed a longer waiting period formed fundamental analysis strategies (Lebaron et al., 1999).

Textual Representation

Those that have attempted of analyzing financial news articles whether in this research paper or others have used a variety of methods. The de facto standard being the Bag of Words approach. This is a modification of apply a vector representation where article terms are indexed and weighted, assigning the importance to determiners and prepositions which have little contribution to the overall meaning of the article. The Bag of Words approach finds solutions to these problems by removing empty stop-words from the article with the remaining terms being used as the textual representation. A second method is called Noun Phrasing which builds upon the Bag of Words approach using a subset of terms as features

(Moldovan et al., 2003), which can address issues related to article scaling while still encompassing the important concepts of an article (Tolle and Chen, 2000). A third representation is called Named Entities, which is a technique that builds upon Noun Phrases by lexical semantic/syntactic tagging where nouns and noun phrases can be classified under predetermined categories (Sekine and Nobata, 2003). This contrasts with using a differential approach, where concepts can be determined using a distributional analysis (Moigno et al., 2002). More on the three and how they are implemented in this research project in subsequent sections.

Machine Learning Algorithms

There are a variety of machine learning algorithms that researchers use, most starting off with a technical analysis of previous prices at appropriate time frames and performing linear regression analysis to determine the textual keywords. Bag of words analysis with keywords such as 'earnings' and 'loss' often helped the researchers predict outcomes based on classes such as up, down and unchanged. A summary table can be seen below that illustrates the variety of machine learning techniques that are used in this area of research.

Algorithm	**Classification**	**Source Material**	**Examples**
Genetic Algorithm	2 categories	Undisclosed number of chatroom posting	Thomas & Sycara, 2002
Naïve Bayesian	3 categories	Over 5,000 articles borrowed from Lavrenko	Gidofalvi et al. 2001
	5 categories	38,469 articles	Lavrenko et al. 2000
	5 categories	6,239 articles	Seo et al. 2002
SVM	3 categories	About 350,000 articles	Fung et al. 2002
	3 categories	6,602 articles	Mittermayer, 2004

Table 1: Prior Algorithm Research

As can be seen in the table, a wide variety of algorithms are used in this area of research. The conductors of this research also notices that most common approach is to classify the predicted stock movements into a set of classification categories and not discrete price predictions. While considering which algorithms to use, they found flaws as some of the sources of information such as discussion board were susceptible to bias and noise. Flaws in Naïve Bayesian methods because using a weighted vector of keywords can unintentionally attach weight to a casually-mentioned security. So they chose to use the support vector machine approach analysis of textual news articles to perform discrete prediction from numeric trends and not simply a binary classification in two predefined categories answering questions such as; *will this article cause the stock price to increase/decrease?* (Schumaker and Chen, 2009).

Financial News Article Sources

There are a vast amount of textual data arising from two sources available to be used from trading and analysis; company or independetly generated. Company generated sources such as quarterly and annual reports can provide a rich linguistic structure that if properly read can indicate how the company will perform in the future (Kloptchenko et al., 2004). While, Independent sources such as analyst recommendations, news outlets, and wire services can provide a more balanced view of the company and have a lesser potential to bias news reports (Schumaker and Chen, 2009) . News outlets such Bloomberg, Business Wire, CNN Financial News, Dow Jones, Financial Times, Forbes, Reuters, and the Wall Street Journal are all sources (Cho, 1999). Furthermore, Stock Quotes provide another form of information and can be divided into various increments of time from minutes to days, however, one minute increments provide sufficient granularity for machine learning (Schumaker and Chen, 2009). The table below provides a summary of textual data available for Analysis.

Textual Source	Types	Examples	Description
Company Generated Sources	SEC Reports	8K 10K	Reports on significant changes Annual reports
	Analyst Created	Recommendations Stock Alerts	Buy/Hold/Sell assessments Alerts for share prices
Independently Generated Sources	News Outlets	Financial Times Wall Street Journal	Financial News Stories Financial News Stories
	News Wire	PRNewsWire Yahoo Finance	Breaking financial news articles 45 financial news wire sources
	Discussion Boards	The Motley Fool	Forum to share stock-related information

Table 2: Examples of textual financial data

3 Problem Statement

Combining regression based methods and textual representation techniques to a supervised machine learning algorithm such as SVM was not a commonly used method in research and was found to possibly lead to a trained system with discrete numeric output as was the goal of this research project. In order to achieve the mentioned objective, research questions can be divided into 2 separate parts:

- How effective is the prediction of discrete stock price values using textual financial news articles?
- Which combination of textual analysis techniques is most valuable in stock price prediction?

4 System design

In order to answer the questions that were raised in the previous section a system called Arizona Financial Text System (AZFinText), illustrated in the figure 2, was build.

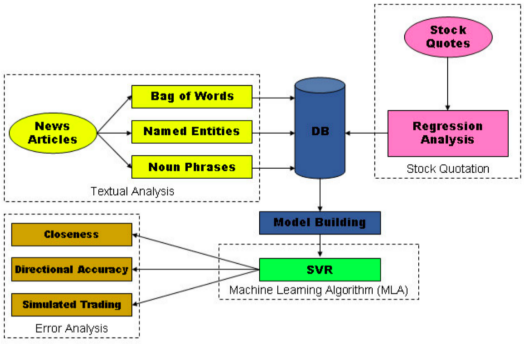

Figure 2: AZFinText system design

Each part of the system is can be seen in the different parts of the figure 2 scheme in detail. First of all, we will be focused on the creation of a consistent database, beginning with the Textual analysis. As one could observe on the figure 2, a Support Vector Machine (or SVM) was created in order to build the model using supervised learning techniques.

Textual analysis techniques

To limit the size of the feature space, just the words that appear three or more times in a document had been taken into account.

- Bag of Words: Is a method that uses language processing to represent documents without taking into account the order of the words. Since every document can be represented with one of those "bags" we can easily process it (lemmatization, stemming, deletion of stop words...) with the help of dictionaries.
- Noun Phrases: Phrases that have a noun (or indefinite pronoun) as its nucleus.
- Named Entities: Real-world object, such as persons, locations, organizations.. that can be denoted with a proper name.

With the help of a lexicon, seven different categories were tagged to some words in order to determine their importance respect our goal, the prediction of stock markets. The categories were ate, location, money, organization, percentage, person and time.

Regression analysis

In order to perform the regression analysis, stock quotes were gathered on a per minute bases for each stock. When a news article was released, the goal was to estimate the stock price 20 minutes ahead. To perform this, a linear regression with the quotation data of the 60 minutes prior to the article release was done. In this very naive approach, one could already see that if the EMH stands, the regression must be meaningless since the stocks are supposed to react instantaneously with the information. From the point of view of the RW view of the market, this regression is also inconsistent since the changes of the stock prices are supposed to be Markovian and unpredictable. The clue of this regression is to quickly predict the stock prices just 20min ahead, so this model is supposed to be somewhere on the middle of both points of views.

Support Vector Machine

A support vector machine (or SVM) is a supervised learning model that is used for classification and regression analysis and it has associated learning algorithms. These learning algorithms consist on the construction of an hyper plane of high or infinite-dimensional space (finite in our case) and, more specifically, trying to find the function that divides the hyper-space ensuring the largest distance to the nearest training-data points of any class.

Types of models

A total of four types of model were build in order to not just obtain a final good behaving model but to know the improvements of every step and know if was worth it and how can it be improved.

- **Regress** This model uses only linear regression from the previous 60 min of the article release moment to predict the price 20 min ahead.
- **M1** This model uses extracted article terms in order to perform the textual analysis but it does not use information about the stock prices at any moment.
- **M2** This model uses the textual analysis of the article, like the M1 model, but it also takes into account the stock price of the moment when the article was released.
- **M3** This model is a combination of all of the above, since it does both the regression and the textual analysis of the article.

5 Experiment design

To know how well the stock market prediction system that was developed works and how well it reacts to real data, is necessary to test it. Then, analyzing the results obtained we will be able to see how good is our engine and if it is worth using it to invest in a real market. The easiest way to test it is designing an experiment with data from the past, of which we know the evolution of the prices of the stocks after the prediction. Therefore, the results obtained can be compared with the real trend of the stock market shares.

The data will be several news articles and stock quotes from 484 of the 500 companies listed on the SP 500 index, some examples of these companies which stocks will take part on the study are Coca-Cola Comp., Ford Motor Comp., McDonalds Corp., HP Inc., Bank of America, Intel Comp. among others. Notice that our method does not analyze mergers and acquisitions, so it implies some small error that will not be considered, but affect at least 2% of the stocks tracked.

The period in study is from October 26th to November 28th, 2005, and it allows us to have enough data for regression trends and future estimation purposes. It have been added a limitation to this data to avoid more than one possible response of the system, this limitation is the fact of using only one article when more than one matched in time and subject.

More specifically, the data is composed by 9211 possible news articles to use and more than 10 millions of stock quotes over the period in study. Using the three different textual representation (bag of word, noun phrases and named entities), and retaining only these terms that appear three or more times in one news article, the data gets filtered into terms of interest, these terms will give us the information to know if the news articles are providing positive or negative to the investors, and so, if the price of the stock quotes will go up or down.

The evaluation of the model will be done using three different metric:

- In **Closeness metric**, it will be computed the difference between the price the machine have predicted and the real stock price it achieved, easily to find as we are working with past data. This difference will be calculate as a mean squared error (MSE).
- **Directional accuracy** measure the direction of the stock price, if it goes up or down. This also could be measured with the closeness metric, but when the changes are small this first metric could be low although the direction of prediction is wrong.
- The last metric is the most applied one, here it is used a **Simulated Trading Engine** which after evaluating each news article, it will buy the stock if the predicted price after 20 min of buying is greater than or equal to 101% of the price when the stock is bought. After this 20min the stocks are sold. If the final among of money is greater than the initial one, we could conclude that it would have worked, at least for the period of time analyzed and the S&P 500 index.

6 Results

In order to know the effectiveness of the stock market prediction system, it is necessary to test all the possible combinations of models with each of the text representations. With this we will have 12 different global models, as there are 4 different models (M1, M2, M3 and a regressed estimate without any use of the news articles) and 3 different textual analysis techniques (bag of words, noun phrases and named entities). Notice that the regressed estimation does not need any text representation, as it uses only 20min of stock price to make the estimation, but still it have to be done separately as we articles chosen for each text representation are different . Each combination can be evaluated with the three metrics explained before. These will allow us to have an image of what combinations are the best, and which are useless.

Closeness metric

The MSE is computed for all the 12 cases, the results depend on the price stock used, so the value alone will not give a lot of information. Despite this, as all the cases are applied to the same data, only comparing the results we will be able to see which combination of methods are the best for this metric.

Trading System	Regress	M1	M2	M3	Average*
Bag of Words	0.07279	930.87	0.04422	0.12605	0.08102
Noun Phrases	0.07279	863.50	0.04887	0.17944	0.10037
Named Entities	0.07065	741.83	0.03407	0.07711	0.06061
Average	0.07271	848.15	0.04261	0.12893	

Table 3: Closeness results for all the different models and text representations.

In the table, it can be seen first that M1 method does not achieve any kind of approximation to the final stock price, this is because it do not have any reference value to know the price of this stocks, and so its very difficult to guess how will evolve the price, if we do not know the price at first instance. For the other three methods, the results are more hopeful. The difference between the predicted price and the real price is small in all the cases, being the best the M2 method, where we use the articles information and the price at the star of the prediction. Strangely, the M3 method, which we expected to be the best one, as it combines regression and articles information obtain worst results than the regress alone. For the text representations, the differences are smaller, but we can say due to the value of average* (an average of the all the methods except the M1) that Named entities are the best option. Also for the M2 method, we achieve the best result when combining with named entities, a MSE of 0.03407.

Directional accuracy

Here we measure the direction of the stock price, so the exact price at the beginning should have no influence, because we just want to know if it goes up or down, and not the magnitude of this change. Doing this randomly we expect to find a a 50% accuracy, so any value higher than this will be a good prediction.

Trading System	Regress	M1	M2	M3	Average
Bag of Words	54.8%	52.4%	57.0%	57.0%	55.3%
Noun Phrases	54.8%	56.4%	58.0%	56.9%	56.5%
Named Entities	54.2%	55.0%	56.4%	56.7%	55.6%
Average	54.6%	54.6%	57.1%	56.9%	

Table 4: Directional accuracy for all the different models and text representations.

Now, the model M1 can be considered as the others, because as it have been said, the value of the stock is irrelevant. It can be seen that we also achieve the best prediction with the M2 model, but now the text representation that helps most is the noun phrases, before, it was the named entities. With this, to predict the direction of the price stocks, the best combination will be the M2 model with the noun phrases. Besides, all the combinations obtain good results, with accuracy over the 54% in all cases. Talking about the text representations, on the contrary of closeness results, here we find out that the best is the noun phrases representation.

Simulated Trading Engine

The best way to see how the system works, and if it is possible to obtain some benefit from this predictor, is to simulate inversions using the predictions done. The results are more clarifying than before as the

amount of money that can be earned by using this engine, and this is the main objective when buying and selling on the stock market. The benefits are expressed in percentage respect to the initial inversion.

Trading System	Regress	M1	M2	M3	Average
Bag of Words	−1.81%	−0.34%	1.59%	0.98%	0.42%
Noun Phrases	−1.81%	0.62%	2.57%	1.17%	0.64%
Named Entities	−2.26%	−0.47%	2.02%	2.97%	0.57%
Average	−1.95%	−0.05%	2.06%	1.67%	

Table 5: Simulated trading benefits for all the different models and text representations.

It can be seen some hopeful results in table 3. Using method M2 and M3, some benefit is obtained, independently of the text representation we use. The most benefit is when combining the M3 method with named entities with a not negligible 2.97%, and the second one with more benefit is a completely different combination, method M2 with noun phrases, which achieve a 2.57%. The results obtained by the regress are unexpected, as the negativeness of the variation is too high to be considered as a coincidence in case regress does not predict anything. An explication could be that when an article appears, it is expect to be a reaction in the stock price, and regress does not notice it. This reaction is usually negative, and so the price decrease rapidly, making us lose money if this method is used. For the M1 method, it can be seen that using articles alone is insufficient.

All the metrics find that the best method is the M2, also in simulated trading engine with an average benefit of 2.06%. For the text representation the results are not as clear, while in closeness metric the best results are found with named entities, the other two metric match in that the best choose is noun phrases. Strangely, named entities give the worst result for closeness metric. What we agree is that the bag of words, although not getting bad results compared to the other representations, is the representation that gets worst results.

7 Discussion of results and conclusions

The main purpose of this paper was to answer the two questions that were raised in section 3. Both questions can now be answered with the results above.

The first asked how effective is the developed method, this can be seen in table 3, there it is seen that the effectiveness of the combination of stock price and textual financial news articles is good to obtain some benefit. The results give us enough confidence to say that this method could be used to inversions as the data gets large enough, and so the possible probabilistic deviation should be smaller than the results achieved. Although this, the effectiveness is smaller as it is needed a lot of data to obtain some kind of benefit.

The second question asked is not as easy to answer as the previous one. If we look at the average from all the models, the most valuable textual analysis technique is the noun phases, but it is the worst at predicting the closeness. The named entities also are good as expected, they achieve the second best performance in directional accuracy and simulated trading engine, and are the best in closeness measure. What is more clear is that bag of words textual analysis is the worst one, as expected, probably due to the high quantities of noise words that contain the articles used. To be more sure about this, we can only take the M2 model, with this we avoid taking into account some aberrations that could produce regress and M1 methods, as both are not very accurate. The results are more or less the same as for the average of all models, for closeness metric the named entities are the best choice, and for both direction accuracy and simulated trading, the best is noun phrases.

8 Future research

Finally, some ideas are posted in order to give direction for a possible future research. The first idea, is to use other machine learning techniques, such as relevance vector regression. This implementation could mean having fewer vectors in classification and better accuracy.

Another idea to improve the model would be expanding the selection of stocks outside of the S&P 500 ones. The S&P is a fairly stable set of companies, so perhaps a more volatile and less tracked companies analysis may provide interesting results. Related to this idea two more ideas come naturally up. The first one is to use a larger dataset, that the authors think would help offset any market biases that are associated with using compressed periods of time, such as the effect of cycles stocks, earning reports, mergers and other wild possibilities. Since the system focuses on the entire S&P 500, it would be also a great idea to be more selective when training. For example, making industry groups or company peer group.

Finally, should be also interesting to test the behaviour of the model now based on the percentage of stock price change instead of fixed stock prices. It could be applied, for example, on penny stocks that tend to have wild fluctuations.

9 Bibliography

Bourgi, S. (2020). Dow dumps while boeing stock plunges to staggering 28 billion dollar loss.

Cho, V. W. (1999). Knowledge discovery from distributed and textual data, in computer science.

Fama, E. F. (2016). The behavior of stock-market prices author (s) :.

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., and Visa, A. (2004). Combining data and text mining techniques for analysing financial reports. *International Journal of Intelligent Systems in Accounting*, 12:29–41.

Korte, G. (2019). Ethiopian airlines flight crash: 157 people dead.

Lebaron, B., Arthur, W. B., and Palmer, R. (1999). Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control*, 23(9-10):1487–1516.

Majaski, C. (2020). The difference between fundamental vs. technical analysis?

Moigno, S., Charlet, J., Bourigault, D., and Degoulet, P. (2002). Terminology extraction from text to build an ontology in surgical intensive care. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 430–4.

Moldovan, D., Pasca, M., Harabagiu, S., and Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst.*, 21(133-154):133–154.

Schumaker, R. P. and Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2).

Sekine, S. and Nobata, C. (2003). Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Tajik, I. (2019). The best fundamental indicators for forex trading.

Tolle, K. and Chen, H.-c. (2000). Comparing noun phrasing techniques for use with medical digital library tools. *JASIS*, 51:352–370.