Applying big data technologies in the financial sector – using sentiment analysis to identify correlations in the stock market

Eszter Katalin Bognár*
Business Informatics M.Sc., Budapest University of Technology and Economics, Hungary

Abstract

The aim of this article is to introduce a system that is capable of collecting and analyzing different types of financial data to support traders in their decision-making. Oracle's Big Data platform Oracle Advanced Analytics was utilized, which extends the Oracle Database with Oracle R, thus providing the opportunity to run embedded R scripts on the database server to speed up data processing. The extract, transform and load (ETL) process was combined with a dictionary-based sentiment analysis module to examine cross-correlation and causality between numerical and textual financial data for a 10 week period. A notable correlation (0.42) was found between daily news sentiment scores and daily stock returns. By applying cross-correlation analysis and Granger causality testing, the results show that the news' impact is incorporated into stock prices rapidly, having the highest correlation on the first day, while the returns' impact on market sentiment is seen only after a few days.

Keywords: *Big Data, finance, Oracle Advanced Analytics, sentiment analysis, R programming.*

1. Introduction

"To make the most of big data, enterprises must evolve their IT infrastructures to handle these new high-volume, high-velocity, high-variety sources of data and integrate them with the pre-existing enterprise data to be analyzed."[1] Financial markets can be viewed as complex, dynamic systems that evolve continuously and interact with the news as well as the economic and political environment. As a side effect, they generate enormous amounts of data and challenge the utilization of the latest Big Data technologies.

This article consists of two main parts. The first is a technical overview of the latest Oracle Big Data technology which was used to create a data analysis system for processing financial market data. In the second part, a sentiment analysis study will be introduced to discover the connection between market sentiment and stock returns.

2. Technology overview

In the following section, the Oracle Big Data options' [2] [3] [4] applicability in designing a financial data processing system will be reviewed.

2.1. Oracle Advanced Analytics

Oracle Advanced Analytics (OAA) is an option to the Oracle Database Enterprise Edition and offers in database analytics. The main advantage of this option is that it brings the algorithms close to the data. In contrast to traditional data analysis platforms where it is required to transfer the data from the database to separate analytical and statistical engines, using Oracle Advanced Analytics enables to use the algorithms where the data are stored thus eliminating the overhead of moving the data back and forth. This gives a better, simpler, more scalable architecture for delivering better decisions and deeper insights using predictive analytics. This platform is ideal for processing and forecasting financial time series data and news, where real-time processing is a requirement. It also lowers the total cost of ownership, as it eliminates the need for separate analytical servers.

The Oracle Advanced Analytics option comprises both the Oracle Data Mining (ODM) and Oracle R Enterprise (ORE) components. It offers a wide range of data mining and statistical algorithms. With Oracle Advanced Analytics, it is easy to discover hidden patterns in massive volumes of data, helping users to reach new insights, form predictions and quickly apply results. The native SQL analytics and model building, as well as the embedded R execution, are made parallel so that the processing of high volume data is much faster than

[*] corresponding author email: bgeszti@gmail.com

with traditional architectures. OAA also offers powerful visualization tools, as the results of the analysis can be accessed via different kinds of UIs like Oracle Business Intelligence Enterprise Edition[1] or web database applications created using Oracle Application Express (Apex)[2].

Oracle Advanced Analytics can be accessed through a graphical interface, like that contained in Oracle Data Mining, or via SQL, PL/SQL and R APIs. Database applications have direct access to the OAA database tables, enabling the user to easily visualize results and create dashboards for ad hoc analysis.

2.2. Oracle R

For complex analyses of large datasets, which are usually stored in an Oracle database, it is much faster and easier to do this from "inside" the database, rather than exporting the data into another specialized external format.

Oracle R Enterprise, a component of the Oracle Advanced Analytics option, makes the open source R statistical programming language[3] and environment available for in-database Big Data processing. R has become very famous in recent years; it boasts a global, highly active community of approximately two million users, with new packages being developed daily.

2.3. In-database analytics using Embedded R execution

Oracle R Enterprise lets users run R scripts in the database using embedded R execution. Server-side execution can be done via R or SQL interfaces. The R interface is most often used to test R scripts before utilizing them in database applications. The SQL interface can be used in database applications.

By utilizing server-side resources, server-side execution of scripts helps to eliminate the memory constraints of the R client while also providing the benefits of parallel execution.

3. Sentiment analysis

In recent years, sentiment analysis has become a hot topic among researchers attempting to automatically extract and quantify the opinions and sentiments embedded in news headlines, financial microblogs and Twitter tweets.

Sentiment analysis is itself is a data mining technique, the primary goal being to automatically extract the news' attitude towards the subject of analysis without having to manually read content. The first step of analysis is cleaning up the text in question. News on webpages is generally in HTML format, so HTML tags and irrelevant items must be removed. After converting the text into a clean format, different pre-processing techniques are applied. One of the early-stage techniques is to identify the "units" of the text; these may be words, sentences, phrases or n-grams. Lemmatization or stemming are also used to reduce the number of words by simplifying them to their common root. Additional techniques include removing capital letters, identifying the language of the text, removing stop-words, etc.

The most basic form of sentiment analysis is performed by counting the positive and negative words in the text, then using the resulting proportion to arrive at a sentiment score.

After the selection and preliminary processing of the text units, it is imperative to choose the right dictionary for the sentiment analysis. A given word can express a positive attitude in the field of finance but a negative attitude in some other field. For example, if we look at the sentiment of the word "rise" in the financial sector, it usually means a positive thing: rise of prices. If we examine the same word in the context of healthcare, it may have negative connotation, such as rise of blood pressure. For this reason, it is imperative to select the sentiment dictionary that best fits the domain of the analysis.

Besides the application of sentiment dictionaries, statistical methods that identify the sentiment of a given article can be employed. The most popular statistics-based solutions incorporate the naïve Bayes or Maximum Entropy classifiers or Support Vector Machines, but neural networks can also be used for this task. The

[1] https://www.oracle.com/solutions/business-analytics/business-intelligence/index.html
[2] https://apex.oracle.com/en/
[3] https://www.r-project.org/

implementation can be done by using either supervised or unsupervised learning methods. For the supervised learning, a training set consisting of articles and their sentiment is required. Most of the time, the training set is created manually with the contributions of experts in the researched field. The shortcomings of this method are the amount of time needed to classify the news in the training set and the subjective nature of sentiment; different people may express a variety of opinions regarding the same article. In unsupervised learning, the stock prices are usually used to train the classifier. If stock prices go up, the articles for the day should express positive sentiment, and vice versa if stock prices fall.

4. Previous studies

One influential person in the early years of sentiment analysis is the Hungarian researcher, Győző Gidófalvi. His work [5] formed the basis of many later studies in the field. Gidófalvi applied a naïve Bayesian text classifier to 5500 financial news articles covering 12 stocks. The change in stock price and the β-value (used to measure the volatility of stock) were fed into this classifier, which then grouped the articles into three categories. Although the prediction power of the classifier was low, Gidófalvi recognized a strong correlation between the news articles and the stock price movements.

A framework called AZFin Text (Arizona Financial Text System) was designed by Robert P. Schumaker and Hsinchun Chen [6][6]. In their research, the prediction power of daily news was tested using three textual methods: Bag of Words, None Phrase and Named Entities. They examined 9,211 financial news articles and 10,259,042 stock quotes during a five week period and tried to predict the stock price 20 minutes after the news articles were released. They achieved 57.1% accuracy in predicting the direction of future stock price movement using the stock price at the time of an article's release.

The work of Anurag Nagar and Michael Hahsler [7] is also notable. They used the open source R language for the analysis; more specifically, a dictionary-based sentiment analysis approach including words from the MPQL dictionary, combined with words built into the tm.plugin.webmining R package that is widely used for sentiment analysis. Although their paper omits the detailed, number-based evaluation of the results, visualization of their analysis shows a strong correlation between news sentiment and stock prices.

Koppel and Shtrimberg [8] investigated 12,000 articles about stocks in the S&P500 index between year 2000 and 2002. They employed support vector machines and naïve Bayes classifiers to predict the change in daily stock prices. The authors selected for all words that appeared at least sixty times in the corpus and eliminated common stop words. They also introduced two complex labelling methods for the financial news based on the date of the article publication and the gap between the daily opening and closing prices. They concluded that negative price movements are more easily predicted than positive ones.

Nuno Oliveira, Paulo Cortez and Nelson Areal [9][9] analyzed data from StockTwits, a microblogging website dedicated entirely to the stock market. This platform is similar to Twitter, in that the length of messages are limited to 140 characters. The team investigated five companies and the S&P500 index with three measurements: stock return, volatility and trading volume. They used a robust regression model to predict stock prices and the RSME and MAPE indicators to measure error. Their analysis was also performed using the R programming language. In contrast with previous studies, they found that there is no strong correlation between stock returns and sentiment indicators or between posting volume and volatility.

Nearly all of the aforementioned studies propose that there is a correlation between sentiment in different textual sources and stock price movements. As the amount of textual information become larger and larger, there is growing need to interpret and incorporate it into stock price predictions.

5. Sentiment dictionaries

The dictionary-based approach is the easiest method for extracting sentiment from text documents. The basic premise is to look at the text of articles as a "bag of words." Meaning, the order and grammatical structure of the words does not matter. From this collection of words, a document-term vector can be constructed that represents each word along with its frequency. Next, the words are matched to entries in the selected dictionary where words and their sentiments are stored. The potential shortcomings of this method are in the selection of the right dictionary and in the methods used to weigh different words. In our analysis, the Harvard Inquirer[4] and The Bing Liu dictionaries[5] were chosen.

The General Inquirer (GI) uses words from the Harvard IV-4 dictionaries or the DICTION dictionary developed by Roderick Hart.

The GI lexicon contains approximately 12,000 words grouped in 184 sentiment categories such as negative, positive, strong, weak, power, pain, etc. Thus, this dictionary not only contains the basic positive and negative sentiments, but a variety of other associations.

The Bing Liu Lexicon contains approximately 6,800 English words categorized as either negative or positive. It also contains mistyped and slang words to extend the applicability of the lexicon.

6. System design

6.1. Data sources

My experiment was conducted using financial data for Apple (AAPL) covering a 10 week period from 28/09/2015 to 04/12/2015. Only the New York Stock Exchange trading days were considered, thus weekends and 26/11/2015 (market holiday due Thanksgiving Day) are excluded from the dataset. The data were stored in Oracle Database tables.

Daily adjusted closing prices were downloaded from Yahoo Finance, and financial news scraped from Google News RSS Feeds with an average of 34 postings per day. Regarding the daily news, only those articles published between 7:30 AM and 16:30 PM were considered, which covers the stock market trading hours plus two additional hours before the market opens. This time window was selected to exclude news which could not affect daily stock prices, on account of its after-hours publication.

6.2. Pre-processing of financial news

The following pre-processing steps were performed on the financial news in the database using the *XML, RCurl, boilerpipeR, tm.plugin.webmining, NLP* and *openNLP* R packages:

- **Filtering content**: Article content can vary. Some articles focus only on the examined company, in this case Apple, while others contain superfluous information regarding other companies. To avoid irrelevant content, each article was broken into its composite sentences and only those containing the words “appl” or “apple” were extracted.
- **Converting the text to lowercase**: Text typical of a news article contains both upper and lower case characters, which can prove cumbersome when searching for words. Prior to the analysis, all text was converted to lowercase, making a given word identical in appearance wherever it occurs.
- **Removing numbers, punctuations**: Numbers, punctuation and special characters were removed.
- **Removing stopwords**: In every text, there are a lot of common but uninteresting words (a, and, also, the, etc.). Stopwords are frequent by nature, and will confound the analysis if they remain in the text. To avoid this, common English stopwords were removed from the articles.
- **Creating a document term matrix**: A core step of text analysis is the creation of the document term matrix. This is an n*m matrix, where n refers to the number of documents and m refers to the number of terms. Each value in the matrix is a frequency number that shows how many times the term appears in the document. This matrix can be very sparse. The terms were sorted based on their frequencies in decreasing order, and only those with frequencies above a given minimum threshold were kept.

[4] http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html

[5] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

6.3. News sentiment scoring using sentiment dictionaries

After pre-processing, the documents were scored according to sentiment. The scoring technique was based on the method of Johannes Tines[6].

Each of the documents was broken into sentences, and filtered for company-specific content. For each of the remaining sentences, a document term matrix was created, and the positive and negative terms counted. For locating the positive and negative terms within the sentences, two different sentiment dictionaries were used (Harvard General Inquirer and the Bing Liu opinion lexicon). For each sentences in the document, the sentence score was calculated by subtracting the negative term count from the positive term count (1) below.

$$sentenceScore = count(posTerms) - count(negTerms) \tag{1}$$

After obtaining these sentiment scores, each sentence was marked as 1, -1 and zero (positive, negative and neutral), by taking the sign of the sentence score (2).

$$sentenceSign = sign(sentenceScore) \tag{2}$$

Finally, the document score for a given article was calculated from the sentence signs by taking the proportion of the positive and the sum of the positive and negative sentence signs (3).

$$documentScore = \frac{sum(sentenceSign == 1)}{sum(sentenceSign\ != 0)} \tag{3}$$

Using this scoring technique, each documents received a sentiment value between 0 and 1.

Since many documents are published within a single day, an overall daily sentiment score can also be calculated from the daily document scores, by taking the sum of document scores greater than 0.5 divided by the sum of document scores greater than 0 (4).

$$dailySentScore = \frac{sum(documentScore > 0.5)}{sum(documentScore\ != 0)} \tag{4}$$

Using the above technique, news sentiments could be quantified and serve as the basis for comparisons with numerical financial data. To perform the analysis, the data were stored in an Oracle Database, with the sentiment calculation and further analysis conducted by Oracle in-database analytics with Oracle embedded R execution.

Fig. 1 shows the design of the financial data processing system.

[6] https://www.linkedin.com/pulse/20141109035942-34768479-r-sentiment-scoring-hsbc-w-harvard-general-inquirer

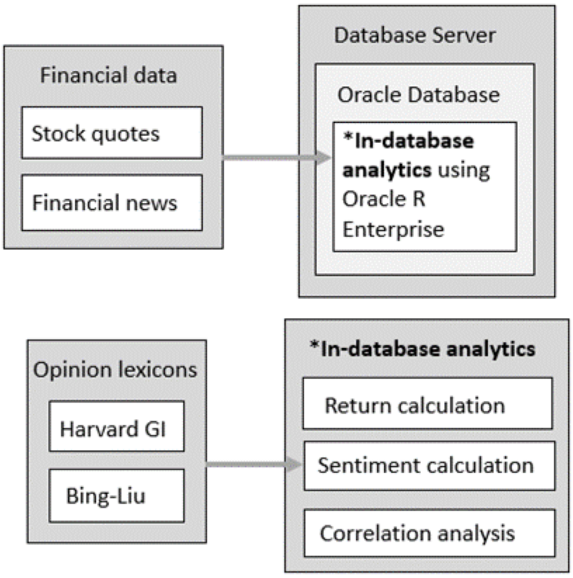


Fig. 1. System design

7. Results

The combined application of the General Inquirer and Bing Liu dictionary can lead to a better correlation between returns and sentiment scores. To identify the best weighing method, I cycled through all of the possible combinations of weights with 0.1 steps. The resulting table (**Table** 1) shows the best correlation in the case of 0.6 and 0.4 weights for the General Inquirer and Bing Liu lexicons, with a 0.423 correlation coefficient.

Table 1. Correlation with Different Dictionary Weights

GI score	Bing-Liu score	Combined score	GI weight	Bing-Liu weight
0.340	0.311	0.331	0.1	0.9
0.340	0.311	0.353	0.2	0.8
0.340	0.311	0.376	0.3	0.7
0.340	0.311	0.397	0.4	0.6
0.340	0.311	0.414	0.5	0.5
0.340	0.311	0.423	0.6	0.4
0.340	0.311	0.419	0.7	0.3
0.340	0.311	0.402	0.8	0.2
0.340	0.311	0.375	0.9	0.1
0.340	0.311	0.331	0.1	0.9

A visual comparison of the sentiment score and daily return plots (**Fig. 2**) reveals a possible correlation. The general up and down course of the two time series are quite similar. It is also remarkable that some lags can be identified. A significant decrease in the returns of Apple on 26/10/2015 was followed with a negative market sentiment on the next day.

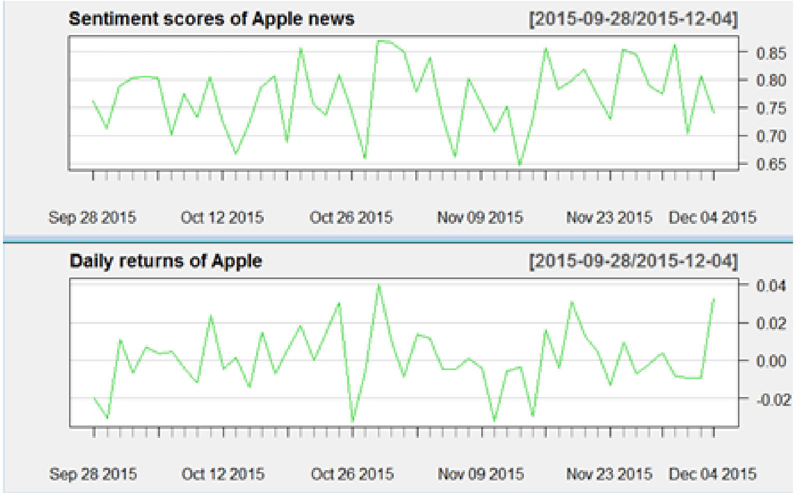


Fig. 2. Daily stock returns and sentiment scores

In some time intervals, the direction of time lag is reversed; that is, the changes in daily returns appear to follow the changes in market sentiment. To clarify this, a cross correlation analysis was taken on the two time series to identify the possible lags between the data. **Fig. 3** shows the cross-correlation between the returns and sentiment scores.

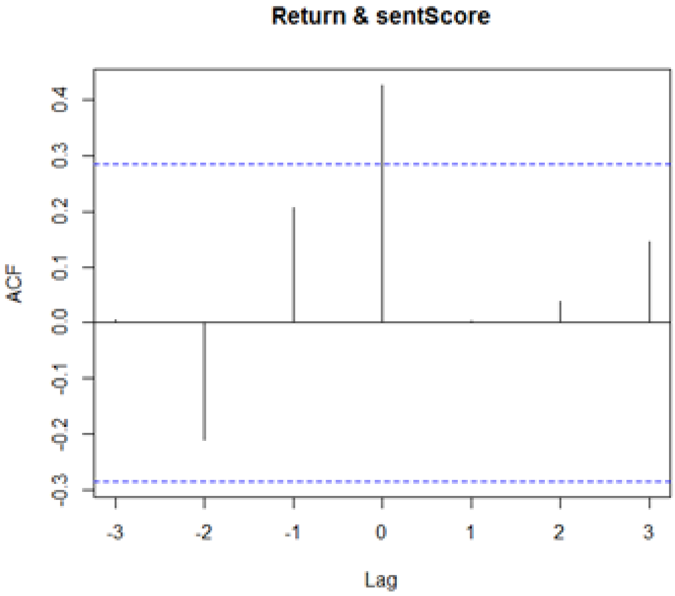
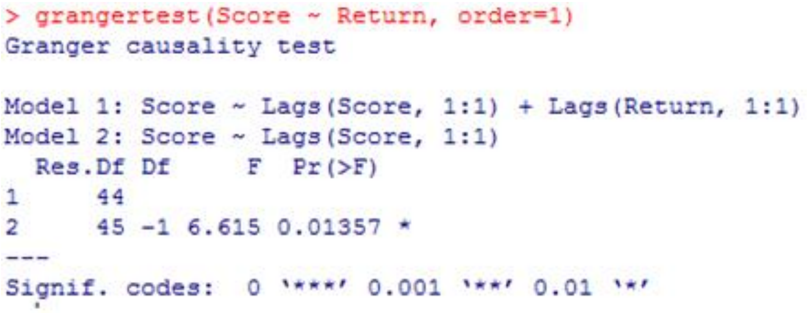


Fig. 3. Cross-correlation

Examining this correlation plot, it can clearly be seen that the highest correlation between the returns and sentiment scores is on the first day, meaning that the news' influence works quickly on stock prices. It is interesting that a weaker but still notable correlation can be seen on lag -1 and -2, indicating that returns may affect sentiment score values a few day later. Therefore, the market sentiment is a reaction to market prices, and does not have predictive power for future stock returns.

Another way to identify the relationship between the two time series is the usage of the Granger causality test. The main question is whether the future values of a time series are predictable using prior values of another time series. This is tested by using the lags of one series to model changes in a second series.

Based on the results of the cross correlation analysis, I suggest that the return values at time *T-1* can predict the sentiment scores at time *T*. **Fig. 4** shows the result of the Granger test that indicates that returns are granger cause sentiment scores with a significant (0.01357) p value.



```
> grangertest(Score ~ Return, order=1)
Granger causality test

Model 1: Score ~ Lags(Score, 1:1) + Lags(Return, 1:1)
Model 2: Score ~ Lags(Score, 1:1)
  Res.Df Df     F  Pr(>F)
1     44
2     45 -1 6.615 0.01357 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
```

Fig. 4. Granger causality

8. Conclusions

Although the predictive power of the sentiment scores cannot be proven, it is apparent that the sentiment of the financial news is correlated with returns. The selection of sentiment scoring method and data sources are important for obtaining statistically meaningful results. Both the General Inquirer and the Bing Liu lexicons were appropriate for identifying stock market direction, with the best performance achieved using their weighted combination. Both the cross correlation and Granger tests demonstrated that the effect of the news on returns is fast-acting, even appearing on the same day. On the other hand, the news is influenced by stock returns only after a few days lag.

The analysis could be improved by incorporating more sophisticated semantic analysis of the textual data and by using more domain-specific sentiment dictionaries.

With the detailed review and appropriate selection of the latest Oracle Big Data technologies, the data processing system outlined in this article can form the basis of further real time analysis of structured and unstructured financial data in a fast, efficient and cost-effective manner. This article hopes to introduce and encourage the use of these new technologies in the financial sector.

References

[1] Oracle: "*Big Data for the Enterprise*", Oracle White Paper, June 2013, [Online]. Available: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf [Accessed March 29, 2016].

[2] Charlie Berger: "Big Data Analytics with Oracle Advanced Analytics In-Database Option", 2012, [Online]. Available: http://www.oracle.com/technetwork/database/options/advanced-analytics/oaa12cpreso-1964644.pdf [Accessed March 29, 2016].

[3] Bryan Pottle: "*Introducing Oracle R Enterprise 1.4*", 2014, [Online]. Available: http://www.oracle.com/webfolder/technetwork/tutorials/tutorial/db/ore1.4/part1/presentation_content/external_files/Introducing_Oracle_R_Enterprise_1_4.pdf [Accessed March 29, 2016].

[4] Bryan Pottle: "*Embedded R execution: R Interface*", 2014, [Online]. Available: http://www.oracle.com/webfolder/technetwork/tutorials/tutorial/db/ore1.4/part6/presentation_content/external_files/Embedded_R_Execution.pdf [Accessed March 29, 2016].

[5] Gidofalvi, Győző: "*Using News Articles to Predict Stock Price Movements*", Department of Computer Science and Engineering. University of California, San Diego, 2001, [Online], Available: http://cseweb.ucsd.edu/~elkan/254spring01/gidofalvirep.pdf [Accessed March 29, 2016].

[6] R. P. Schumaker, H. Chen: "*Textual analysis of stock market prediction using breaking financial news: The AZFin Text system*", ACM Transactions on Information Systems (TOIS), Volume 27, Issue 2, Article No. 12., February, 2009.

[7] Anurag Nagar and Michael Hahsler: *"Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams"* in International Conference on Computer Technology and Science (ICCTS 2012), IACSIT Press, Singapore, 2012.

[8] M. Koppel and I. Shtrimberg: **"***Good News or Bad News? Let the Market Decide*" in ***AAAI Spring Symposium on Exploring Attitude and Affect in Text,*** The AAAI Press, Palo Alto CA, pp. 86-88., 2004.

[9] Nuno Oliveira, Paulo Cortez, Nelson Areal: "*On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume*" in 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Heroísmo, Azores, Portugal, pp. 355-365., 2013.