2012 International Conference on Computer Technology and Science (ICCTS 2012)
IPCSIT vol. XX (2012) © (2012) IACSIT Press, Singapore

Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams

Anurag Nagar [1], Michael Hahsler [2]

[1] Computer Science and Engineering, Southern Methodist University, Dallas, TX, USA

[2] Engineering Management, Information and Systems, Southern Methodist University, Dallas, TX, USA

Abstract.
Analyzing stock market trends and sentiment is an interdisciplinary area of research being undertaken by many disciplines such as Finance, Computer Science, Statistics, and Economics. It has been well established that real time news plays a strong role in the movement of stock prices. With the advent of electronic and online news sources, analysts have to deal with enormous amounts of real-time, unstructured streaming data. In this paper, we present an automated text mining based approach to aggregate news stories from diverse sources and create a News Corpus. The Corpus is filtered down to relevant sentences and analyzed using Natural Language Processing (NLP) techniques. A sentiment metric, called NewsSentiment, utilizing the count of positive and negative polarity words is proposed as a measure of the sentiment of the overall news corpus. We have used various open source packages and tools to develop the news collection and aggregation engine as well as the sentiment evaluation engine. Extensive experimentation has been done using news stories about various stocks. The time variation of NewsSentiment shows a very strong correlation with the actual stock price movement. Our proposed metric has many applications in analyzing current news stories and predicting stock trends for specific companies and sectors of the economy.

Keywords: Text Mining, Stock Market, News Sentiment, Stock Trends

1. Introduction

Many key factors that influence stock price of a company or a sector of the economy are also affected by the incoming news articles and feeds [1], [2]. Incoming news can be of various types - such as latest earnings statements, announcement of dividends by a company, information about new products, and trend analysis and prediction by financial experts. Figure 1 shows some of the factors that affect the stock price. Many of these factors are inferred from the incoming news about the company and have a strong influence on the trading patterns and ultimately the stock price itself. Fast growth in micro-blogging sites such as Twitter, have also contributed towards the exponential growth in the amount of text data generated about the financial markets.

Clearly, this is an area in which text and data mining tools and techniques can be employed to provide summary information by extracting important keywords and action phrases from the incoming news stories. More importantly, there is need to find ways to find emotion and sentiment from this corpus of text. This is a hot area of research as indicated by the large number of recent publications, for example [2], [3], [4]. Most of the existing research papers focus on extracting sentiment from a static corpus of text such as financial reports, collection of previous financial news items [5], and analysts' reports. With the advent of online news sites such as Google Finance, Yahoo Finance and MSN Money, financial news is delivered in real-time, streaming format. There is a critical need for tools that can provide an accurate estimate of sentiment from streaming news items.

[1] Corresponding author. Tel.: + 1-832-316-2500

E-mail address: anagar@smu.edu

In this paper, we use the open source software **R** to build a news engine for gathering and aggregating news items from various financial sources. It has the capability to gather news in real time and analyze it for key financial terms and phrases. These phrases are analysed for positive/negative affect and polarity by comparing them with publically available lexicons and dictionaries. Our proposed algorithm filters out irrelevant sentences and noise from news stories and creates a text corpus from only those sentences which are relevant to the stock in question. This approach provides substantially better results than the *"bag of words"* approach taken by most researchers. Additionally, working with focussed sets of sentences allows the algorithm to analyse multiple stories in real time and provide instant sentiment scores, which we refer to as NewsSentiment score . To validate our approach, we will compare our NewsSentiment scores with the movement of the actual stock prices and check for a significant correlation between the two.

The rest of the paper is organized as follows – in section 2, we cover previous work that is closely related to this research. In section 3 we introduce our proposed method and develop a framework for extracting sentiment from streaming news. In section 4, we will present our results and compare it with the actual price movement. Finally, in section 5, some ideas for further improvement of our method will be presented.

2. Related Work

There is a lot of on-going research and analysis in the area of text and news sentiment analysis.

Theussl et al [5] have presented a framework for large scale sentiment analysis using the **R** package *tm*. They use a static text corpus from the New York Times and annotate terms based on their polarity and calculate a *sentiment score*. They also illustrate the use of distributed text mining techniques with the Map-Reduce paradigm. In a related work, the authors present an algorithm to generate sentiment scores from static corpus of newspaper articles and study its relationship with the current economic indices [6]. A drawback of their approach is that they consider the entire text and do not filter down the corpus to only relevant paragraphs or sentences.

In a highly referenced article, Godbole et al [2] present a scheme for extracting sentiment from news articles and blogs. Their approach consists of associating entities with sentiments and aggregating and scoring each entity. They provide a good theoretical background and also a good evaluation of their work. However, they work with a static corpus that has been downloaded at a particular time and is therefore not dynamic. It is also not clear how well their method will work at predicting financial indices and stock prices.

Goonatilake et al [1] present a study which indicates that news items have a strong association with economic indices and oil prices. They evaluate the impact of news items on major indices such as Dow Jones Industrial Index (DJIA), S&P 500, and NASDAQ. They classify news items in four categories and try to measure their impact on the daily movement of stock prices by developing a regression model.

Bean [7] uses keyword tagging on Twitter feeds about airlines satisfaction to score them for polarity and sentiment. This can provide a quick idea of the sentiment prevailing about airlines and their customer satisfaction ratings.

Yu et al [8] present a text mining based framework to determine the sentiment of news articles and illustrate its impact on energy demand. News sentiment is quantified and then presented as a time series and compared with fluctuations in energy demand and prices.

3. Proposed Method

In this section, we will introduce our proposed method and its various steps – gathering and aggregating news, filtering the news corpus, and evaluating the corpus based on the polarity of its words.

3.1. News Aggregation

The first step in our method is to create an on-line news gathering and aggregation engine. It can read current financial news from various sources and then process them. For this, we use the various functionalities available in the open source R framework. Specifically, use the package *tm* and its various plugins such as *tm.plugin.webmining* [9]. We use these in conjunction with Google Finance News Feeds to read news in real time.

3.2. Analyzing and Filtering News Items

Our method starts off by scanning the news items for the stock symbols in question. For example, we expect a news story about Apple Corporation to contain the stock symbol "AAPL" in the body of the news story. In fact, the entire news story may not be very interesting or useful for our analysis. Quite often, financial or stock market news articles are a comparison of several companies or even sectors of the economy. In such cases, it becomes necessary to separate sentences or paragraphs about different companies and work with the filtered sentences.

As an example, Figure 1 shows a snippet of a news article from Bloomberg [10]. It talks about Apple (AAPL) and Coinstar (CSTR) in the same article, yet the article conveys different sentiment for each. A naive approach would simply use word tagging on the entire article, thereby producing erroneous results. The proposed method avoids this by filtering the article to only relevant sentences. It would thus take only the first paragraph into consideration for Apple Corporation (AAPL). Some of the key words and phrases that would be tagged for polarity would be "sank 2.8 percent", "fell", "losing streak" (negative), and "valuable" (positive).

> Apple Inc. (AAPL) sank 2.8 percent, the most since October, to $605.23. After rising to a record on April 9, the most valuable technology company fell for a fourth day in the longest losing streak since December.
>
> Coinstar Inc. (CSTR) surged 7.3 percent to $65.78. The owner of the Redbox movie-rental kiosks said first-quarter sales and profit exceeded its previous projection and lifted its earnings forecast for 2012 to at least $4.40 a share.

Figure 1 A snippet from a news article carrying different sentiments for two companies

For breaking a story down to sentences, we have again used R and various Natural Language Processing (NLP) tools available therein, such as the package NLP [11]. This allows us to look at individual sentences and keep only those that contain the stock symbol. Another approach could be to look at the headlines of news items. This can provide us with a quick summary of the market sentiment. For example, the heading – "Cracks in Recent Leaders: CMG, PCLN, AAPL" indicates negative sentiment for the stock AAPL. Similarly we can also look at specific paragraphs that contain stock symbols.

3.3. Identifying Sentiment Words

Our method can work at different granularity levels – such as sentences, headlines, paragraphs, or even the entire news article. We define an **instance** to be a unit of the granularity level at which the analysis is to be performed. For example, an instance can be one single sentence, or one single headline. Similarly, we define a **corpus** as the collection of relevant instances to the stock in question.

After extracting the relevant instances, we identify the key words contained in them and match them against available sources of positive or negative sentiment terms. We have used the Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon [12] and list of sentiment words from the R package *tm.plugin.tags* [13] to create a database of sentiment carrying words.

3.4. Scoring Text Corpus

News articles are kept in memory in the form of a *document-term matrix* with the instances as rows and the terms as columns. We create scores for the instances based on the following definitions:

Definition 1: An instance is classified as positive if the count of positive words is greater than or equal to the count of negative words. Similarly, an instance is negative if the count of negative words is greater than the count of positive words.

$$s_I = \text{sign}(n_p - n_n) \tag{1}$$

where s_I is the score of the instance, n_p is the number of positive words, n_n is the number of negative words and the sign functions returns the sign of its argument.

For example, the sentence “AAPL has been on a phenomenal rise but investors should be cautious” has 2 positive terms – “phenomenal” and “rise", it also has 1 negative term –“cautious”. So this would be a positive instance. We also define a score for the corpus as follows:

Definition 2: Score of a corpus is defined as the ratio of positive instances to the total number of instances.

$$s_C = \frac{|\{s_I \epsilon\, S \mid s_I > 0\}|}{|S|} \tag{2}$$

where s_C is the corpus score, S is the set of all instance scores in the corpus, and |.| is the cardinality of the set.

For example, if a news story has 8 instances out of a total of 10 filtered instances, the score of the corpus would be 0.8. The score is referred to hereafter as the NewsSentiment score.

4. Results

In this section, we will present the results of our method and also its correlation with the actual stock price movement. The news aggregation engine was run each day and it collected the top 20 most relevant stories for the stock in question. We filtered the news items to relevant sentences and headlines before creating a text corpus from them. These corpuses were scored using the scheme presented in Section 3.

4.1. Analysis for Apple Corporation (AAPL)

Figure 2 below shows a comparison of the score obtained by considering filtered sentences corpus to the stock price variation for Apple Corporation (AAPL) between April 3 and April 18, 2012.

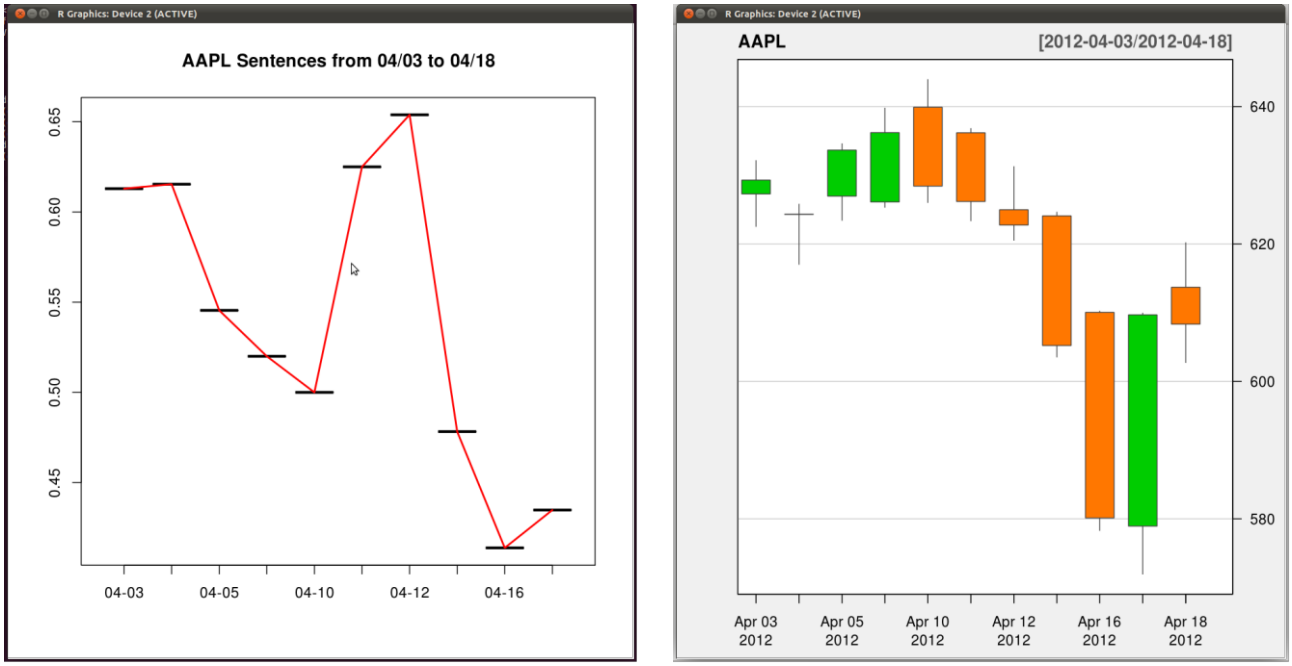

Fig. 2 Comparison of filtered sentences NewsSentiment score and actual stock price for AAPL between April 3 and April 18, 2012

It can be easily seen that there is a strong visual correlation between the NewsSentiment score and the stock price. Another thing to note is that there is a “lag” between the two values. For example, the stock reached its peak on April 10 and the NewsSentiment score reached its peak on April 12. This indicates that the news stories might be descriptive in nature – they may be expressing a negative sentiment after the stock has started falling. This is quite consistent with what financial analysts believe to be true. Our method can thus clearly detect the direction of the market sentiment, which bears a close association with the direction of the actual stock price movement.

4.2. Other Results

Besides the analysis shown for AAPL in section 4.1, we have carried out analysis for various other stocks and also for mutual funds and exchange traded funds. Due to space constraints, we are not including them in this paper. In all cases, NewsSentiment scores had a very strong correlation with the actual stock price variations and can pick up the direction of the market sentiment.

5. Future Work

Our method is able to describe and quantify the market sentiment about stocks quite accurately.. Still, the accuracy can be further improved by considering stock market specific terms into account. For example, "bullish run", "bearish", "short selling" are some terms which are specific to stock market vocabulary but are not part of any normal sentiment lexicons.

As a further step, the authority of the news source can be taken into account while creating the corpus. A higher weight can be given to trusted and authoritative sources. Similarly, if a news items contains too many stock symbols, it can be given a lower weight than one with few symbols. This will allow us to give a higher weight to a focussed news item.

Finally, more resources including social networking and public stock review sites such as stockwits.com can also be taken into account while constructing the news corpus.

6. Acknowledgements

This work was supported in part by the U.S. National Science Foundation under Grant No. IIS-0948893.

7. References

[1] R. Goonatilake and S. Herath. The volatility of the stock market and news. *International Research Journal of Finance and Economics*, 2007, **11**: 53-65.

[2] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. *Proceedings of the International Conference on Weblogs and Social Media* (ICWSM), 2007

[3] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Trends Inf. Retr*. 2008, **2** (1-2): 1-135,

[4] J. Leskovec, L. Backstrom, and J. Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 497-506.

[5] S. Theussl, I. Feinerer, and K. Hornik. 2010. Distributed text mining with tm. In *Proceedings of R Finance* 2010.

[6] P. Hofmarcher, S. Theussl, and K. Hornik,. Do Media Sentiments Reflect Economic Indices? *Chinese Business Review*. 2011, **10**(7): 487-492

[7] J. Bean. R by example: Mining Twitter for consumer attitudes towards airlines. *In Boston Predictive Analytics Meetup Presentation*, 2011.

[8] W.-B. Yu, B.-R. Lea, and B. Guruswamy. A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting. *International Journal of Electronic Business Management*. 2011, **5**(3): 211-224.

[9] M. Annau. tm.plugin.webmining: Retrieve structured, textual data from various web sources. 2012, R package version 0.1/r37. [Online]. Available: http://R-Forge.Rproject.org/projects/sentiment/

[10] R. Nazareth. S&P 500 Caps Biggest Weekly Decline in 2012 on Economy. 2012, [Accessed 15-April-2012]. [Online]. Available: http://www.bloomberg.com/news/2012-04-13/u-s-stock-index-futures-decline-as-china-s-growth-slows.html

[11] I. Feinerer and K. Hornik. openNLP: openNLP Interface, 2010, R package version 0.0-8. [Online]. Available: http://CRAN.R-project.org/package=openNLP

[12] T. Wilson, J. Wiebe, and P. Homann. MPQA Subjectivity Lexicon. 2005, [Accessed 18-April-2012]. [Online]. Available: http://www.cs.pitt.edu/mpqa/subj lexicon.html

[13] S. Theussl. tm.plugin.tags: Text Mining Plug-In: Tag Categories, 2010, R package version 0.0-1.