Predicting Stock Prices from News Articles

Jerry Chen, Aaron Chai, Madhav Goel, Donovan Lieu,
Faazilah Mohamed, David Nahm, Bonnie Wu

The Undergraduate Statistics Association - Project Committee
Fall 2015, Berkeley

December 11, 2015

1 Introduction

The stock market is influenced by a vast variety of sources. Both external factors as well as internal factors move the stock market. From something as small as a press release of a particular firm to something as big as laws passed by the gorvernment, the impact on the stock market is sometimes very visible, prompting many researchers and investors to ask whether the market is predictable. On the contrary, the Efficient Market Hypothesis (EMH) states that the stock price is a reflection of all available sources of information, including news reports, and therefore it is not possible to capitalize on asset mispricing or predict on future asset returns. There are strong and weak forms of the hypothesis, but essentially what it implies is that the market is always efficient. However, criticisms of the theory suggest that this is not always the case. There can still be short-term deviances of pricing from under-researched or unexplored signals. The purpose of this research project is to analyze one of these factors and to test whether the Efficient Market Hypothesis still holds in its context.

2 Motivation

Just like the above mentioned external factors, news articles are one such factor. Many investors consider them to make a decision about investing in a company, potentially affecting the stock prices. Since many articles convey either a positive or negative message, it is reasonable to want to examine the overall sentiment of each text or just the overall sentiment of the first and last paragraphs. We combine these two ideas, stock market impact and sentiment analysis, to analyze news stories from credible sources [1] and to help answer the

[1] We shortlisted news articles written by credible sources only. Credible sources were defined by the amount of readership they had. We also considered news articles that mention the company explicitly. The article could cover any recent move by the company or how a policy

question of whether we can relate this sentiment to the impact on the short-term stock price within the next 3 hours.

3 Methodology

We took a random sample of 30 small cap companies ($300 million - $2 billion in market cap) and pulled 40 articles for each company. We pulled the news articles from Google News, which provided us with reputable news sources. We then stored each articles information, such as when it was published, the headline, and the content of the article, and we moved on to the process of scoring each article as positive or negative.

The scoring process involved breaking down each sentence into positive and negative words and giving specific words different weights. A word that is in the list of positive words that we provide is scored as a +1, while a word that is in the list of negative words is scored as a -1. If one of these words is preceded by a word such as very or another word that amplifies the sentiment of the following word, the weight of the positive or negative word is increased to +1.8 or -1.8. The scoring also takes certain conjunctions into account. For example, if *but* is used in a sentence, the clause following the *but* will be weighted more heavily than the first clause before the *but.* This scoring procedure also applied to other conjunctions such as *however* and *although.* The scoring was done with an R package called **sentimentr** (https://github.com/trinker/sentimentr).

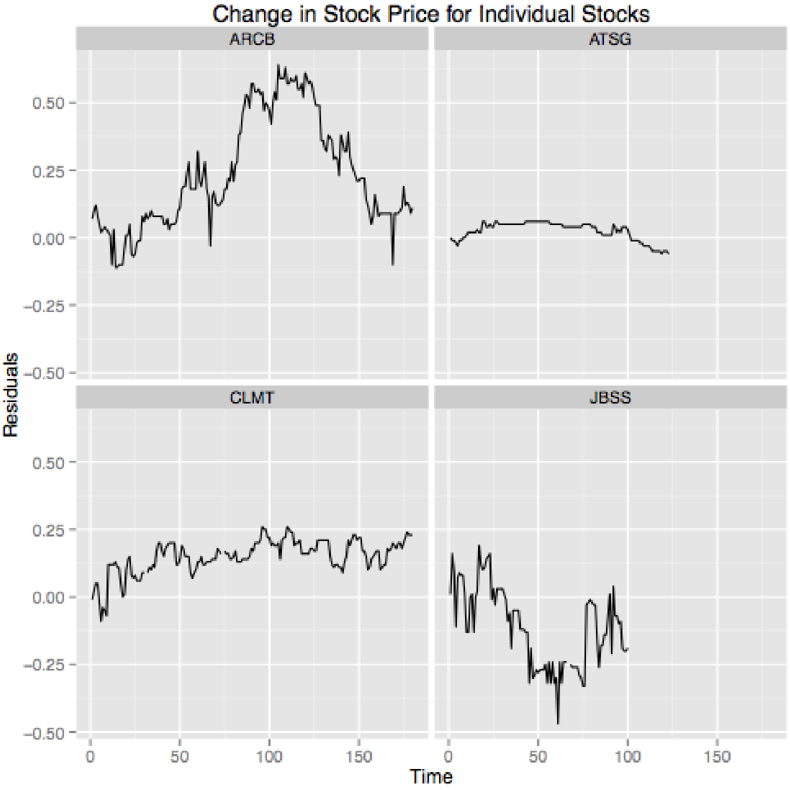
We used this scoring process to score each sentence in an article. We then summed up the scores of all of the sentences in the article to give us the total score for the article. After scoring the article, we either classified it as a positive or negative article. Once we had scored each article, we filtered out the articles that had extremely high or low net scores. We were forced to do this because of the limitations of having to manually extract pricing data for each article. Once we filtered out the articles we wanted, we stored the relevant information (headline of the article, timestamp, symbol, etc.) into a dataframe for a total of 122 articles. We extracted the relevant stock prices for each article from the WRDS database and chose to extract the ask prices 3 days before and 7 days after the publication date for each article.

Each dataset consisted of 10 days worth of data, excluding weekends, leaving about 7 days per article. Each row of the dataset contains the date and time of the measurement, with less than millisecond precision, as well as the relevant ask price, bid price, and the stock ticker. Since the trading hours of the New York Stock Exchange are only from 9:30AM to 4:00PM, the data was split over the different days and measurements outside this time interval were then omitted. Some measurements had a bid price of zero or an ask price of zero; these do not correspond to actual stock quotes, so they were removed. Next, in order to restrict the size of the data set for easier analysis, a single ask price was computed for every minute. This price was taken to be the minimum of the

could affect the company. This was done in order to make sure that it is clear to the investors which company is being affected by the article.

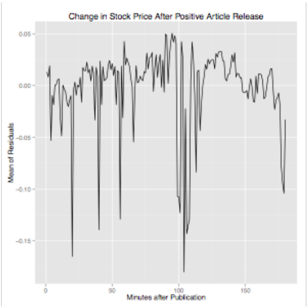
ask price over that interval, intended to represent how little a seller was willing to accept for the given stock, and also to reduce some of the noise inherent in the stock data. Finally this reduced dataset was concatenated over the different days and written to a new file, to be imported for subsequent analysis.

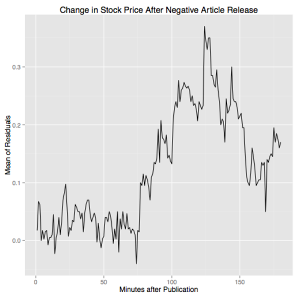
A linear regression model was fitted to the ask price data one hour prior to the articles publication. We then projected this linear model to three hours after the the articles publication date and looked at the residuals. Since there was no noticeable trend, we did the same procedure but did not fit a linear model and just looked at the residuals. We ended up using residuals from not fitting a linear model for all our future analysis. We then plotted the average of the residuals over the positive and negative articles. Also, we ended up only using positive and negative articles that were published during the times the stock exchange were open, and only used the articles in which the stock market was open both an hour before the the publication and three hours after the publication. Due to this, we had only 4 usable negative articles out of 16 and 44 usable positive articles out of 106. The stock price data for all 4 negative articles are shown below.



4 Results

The net price change was analyzed in the short-term three hour range for positive and negative articles. Out of a total of 44 positive articles, 23 articles displayed a net positive price change. A negative net price change resulted from 20 articles, while there was no net change for one article. Out of four negative articles, two had a positive net change while two had a negative net change. Since there appears to be no clear trend in the net price change that results from the type of article, the type of article does not seem to affect the stock price in the short-term time interval. The positive trend in the negative articles could be explained by the fact that the ARCB stock had a very positive trend, and since the average of the negative articles were only taken from the data of 4 articles, it would skew the trend significantly. Displayed are the plots showing average stock price change in the short-term for positive and negative articles.





5 Improvements

Due to the time constraint of the project, the methodology used in this project is just a rough idea of how to determine whether a news article presents a negative image of the related company. There are still potential improvements concerning our methodology.

First, instead of using the built-in dictionary in R, selecting a more comprehensive dictionary of negative words and positive words might enhance the prediction result of whether an article negatively or positively affects a company.

Second, a more robust mechanism for identifying relevant parts (e.g. how many paragraphs in the body of the news and how many sentences in each paragraph are to be selected) of the news articles to do the prediction might enhance the prediction result. For this project, we analyzed the entire news article to see if it is a negative one, but such inclusion will likely contain irrelevant information (e.g. some part of a news article might state negative information of the company in the past (such statements are irrelevant, yet in our mechanism they still affect our prediction), but present positive information in the present).

Third, extracting news articles from different news resources (e.g. Yahoo Finance, Twitter) and combining these articles together to do the prediction might be more accurate (news articles are often subjective per news source, so it is better to select from a variety of samples rather than only extracting information from Google Finance). This can also greatly increase our sample size, allowing for less bias and a more accurate portrayal of the news landscape per stock.

Last, it would be incredibly helpful to sample from all stocks in the market rather than the small-caps. Finding a way to automatically source relevant stock prices from WRDS or any other database can increase our sample size of stocks and offer a more accurate representation of U.S. stock behavior.

6 Conclusion

In the end there were no conclusive evidence of any impact news articles might have on small-cap stock prices in the short-term. As seen in the resulting analysis, a very small sample size sometimes produces counter-intuitive results which is likely not the true behavior of the market. From the angle of positive articles only, there was in general no obvious trend going up or down. Finally, in terms of absolute changes in sign, neither positive nor negative articles have a noticeable influence on the number of stocks whose 3 hour prices were simply above or below the original price. This suggests that we cannot refute the Efficient Market Hypothesis from the scope of this project. However, much can be improved upon for future research - within both sentiment analysis and the design of the project - and the motivation to continue searching for these kinds of opportunities still stands.