Using news to predict stock market trends

Motivation

This is inspired by the Kaggle competition by Two Sigma. The challenge is to use news analytics to predict stock price performance. There is a large amount of market and news data available but the difficulty is finding relevant trends within news data that predict how the stock market will respond. As one would expect, this also means that there may be a large signal-to-noise ratio when it comes to using news analytic data.

A large number of financial institution's primary function is to understand and predict stock market trends. This project would be directly applicable to such institutions. However, the client base for this project goes beyond financial institutions. Any entity that makes use of the stock market can strategically use results from the project to their advantage. By correlating news analytics with stock prices one can make better decisions about one's stock portfolio; one can be better informed as to what stocks to buy and sell and at what times. Ultimately, this project has significance on the global economy as a whole.

Data Description and Wraggling

The data used for this project comes directly from the Kaggle API. The news analytics data was provided by Thomson Reuters. Copyright ©, Thomson Reuters, 2017. While market data is provided by Intrinio. This project uses data from Kaggle resources to predict stock market data from news data. First, two data sets (one for the stock market data and one for the news data) were downloaded from the Kaggle kernel using urlretrieve. A summary of each data set is as follows:

Market Data

· Dataset contains 4,072,956 rows and 17 columns.

· Data contents include date/time stamps for the market data, a company name and identification code, and data pertaining to stock market measurements (e.g. opening and closing prices, volume, returns over 10 day period, etc.)

· There are null values in several columns. In the returnsClosePrevMktres1 and returnsOpenPrevMktres1 columns there are 15,980 null values in each column. In the returnsClosePrevMktres10 and returnsOpenPrevMktres10 there are 93,054 null values in each column. These values are not necessarily needed to model the system and therefore there is little need to change the null values.

News Data

· Dataset contains 9,328,750 rows and 36 columns.

· Data contents include date/time stamps for news segment, the company name and identification code that the news segment is referring to, and details of the news segment (e.g. word count, headline, type of segment, measurement if the segment was positive, negative or neutral toward the company, etc.)

· There are null values in two columns. The headline column has 73,960 null values. The headlineTag column has 6,341,993 null values. Both of these columns will not play a role in the mathematical modeling and there is no logical way of filling them in with alternative values. Hence, the null values are kept in the dataset.

The data sets from Kaggle were already very clean and therefore there was little data cleaning necessary. After downloading and saving the data in csv files these were the steps preformed to clean and join the data sets:

· Each dataset was loaded into python using pd.read_csv(file_name, parse_dates =[1]) so that the date/time was read as a datetime object.

· The null values (see description of each data set above) were all in columns that will not be used in the mathematical analysis. Also, some of the null values (such as headlines) cannot logically be replaced with anything therefore these values were kept null.

· The news data had outliers that are deemed to be statistically realistic and therefore there is no need to adjust for these outliers. The market data had several hundred outliers in several columns that seemed to be a result of miscalculations or incomplete information in calculations. The market data had columns such as opening and closing stock data that had no outliers. It is noted that the outliers are in columns such as returnsOpenNextMktres10 or returnsClosePrevMktres10 which are columns that are calculated from opening price data from the date of the row under observation and data from a date before or after the observation row date. Sometimes a company will appear or disappear from the data set over time; reasons could include bankruptcy, the make or drop from the ranking of companies on the list or the company is newly formed. This means that if a company appears or disappears from the list of companies in the data set there will be an error in calculated some of these values. Hence, the outliers in such cases are not realistic values but rather are calculation errors. It is found that there are 248 outliers in the returnOpenNextMktres10 column; this column is a key value that will be used when modeling. There are two ways one could deal with these outliers. One, the outlier can be replaced with a mean value for that specific company. Two, the entire row can be eliminated completely from the data set. Since there are over 4 million rows of data and only a few hundred outliers it seems reasonable that eliminating the outlier rows would not alter analysis results significantly. Also, there is no basis for changing the value of the outlier therefore it is more logical to remove the data. Hence, the outlier rows were eliminated from the data set.

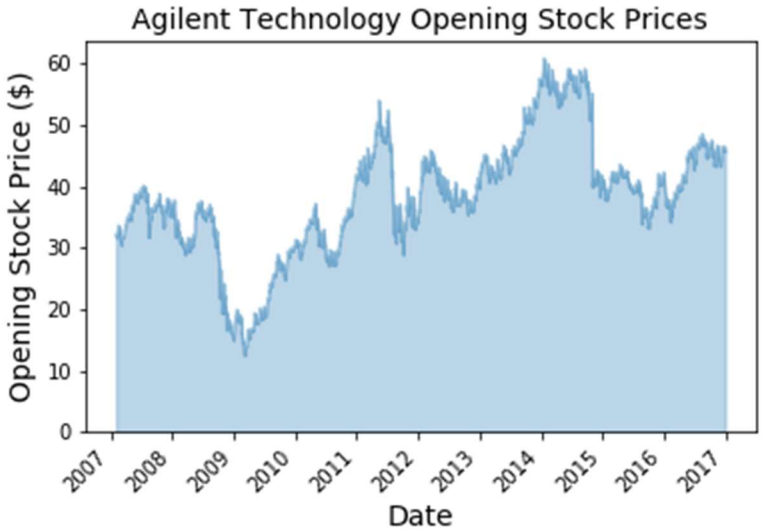
· The last step of data wraggling was to join the two data sets. The logical way to join the data sets was to join them along the company name and the date/time. The challenge is that the news data and the stock data have timestamps at different times during the day. Hence, they will never exactly line up and therefore the data cannot be directly joined using the date/times provided. To join the data effectively we first assume that the data we want to join is only day-to-day and the time of day is irrelevant. Therefore, the datetime variables in the market data and the news data are converted to datetimes with the date only and no timestamp. Then the market and news data were merged along the company names and the date. This resulting dataframe was then saved the a csv file to be used further.

Exploratory Analysis

Stock Prices over Time for Agilent Technologies Inc.

Here an example company is chosen to be used for display of individual company stock data (see graph below). In this case Agilent Technology Inc. is used. Above the opening stock price for the company is shown versus time. As seen in the plot the stock prices tend to look like they

are noisy from day-to-day with long term trends being the predominate trends. For instance, we see that from day-to-day the price may increase or decrease but generally there are no major changes short term. However, in major events such as the 2008 depression a large drop in the stock price is seen over the course of a year. The challenge moving forward is to look a news data to predict both the short term and long term trends. From this graph it is clear that short term trends may be more difficult to predict because of the general noisy nature of day-to-day trading.

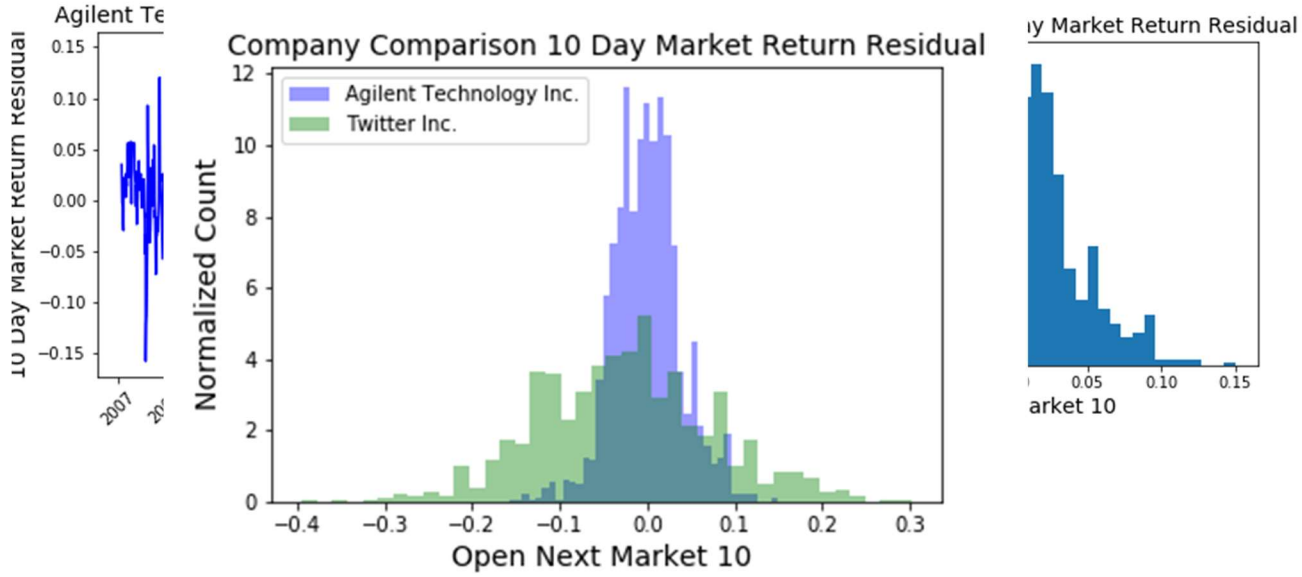


10 Day Market Residualized Return for Agilent Inc.

Here the 10 day market return residual is graphed versus the date. This value is indicative of the amount the stock price changes over the course of 10 days. As seen the value fluctuates wildly from day-to-day. If instead the 10 day market return residual is graphed as a histogram we see that the value looks nearly normal about a value of 0. The distribution actually looks slightly skewed to larger values which makes sense. Because the stock prices ultimately increased over the timeframe in which the data was taken one would expect that there would be more values of the 10 day market return residual that are positive. For this particular stock the overall stock

price increased over the timeframe hence the distribution is skewed to larger values of the residual.

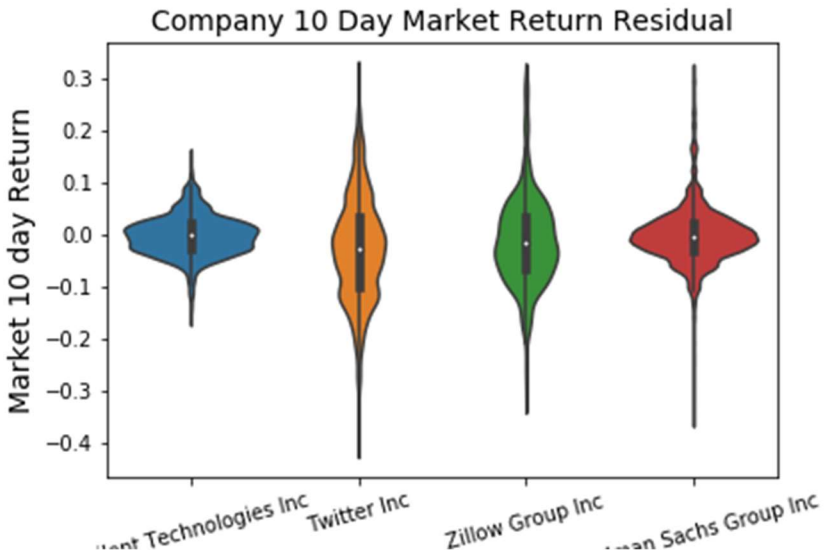
Here is an example of a stock with a very different market trend. Ultimately Twitter Inc. stock dropped over time. When comparing the 10 day market return residual of Twitter to Agilent Technology we see that Twitter has a similar skewed distribution except it is skewed toward negative values. Again, this makes sense because the stock price of Twitter Inc. dropped over time and therefore there should be a tendency for the distribution of the residual to be negative.



10 Day Market Residualized Return for Different Companies

Even better than comparing just two company histograms, we can make violin plots of the 10 day market return residual for several different companies. Notice, many companies are nearly symmetric about zero but are slightly skewed depending on if the company stock had increased or decreased over time. Also, the range of the residual values vastly varies depending on company. This poses an additional challenge because it suggests that some companies are more prone to dramatic changes than others which signifies volatility. Also, if one wanted to use such information to make money off of short trading then the long term trends would be less important. For instance, Twitter ultimately lost stock value over time however there is larger

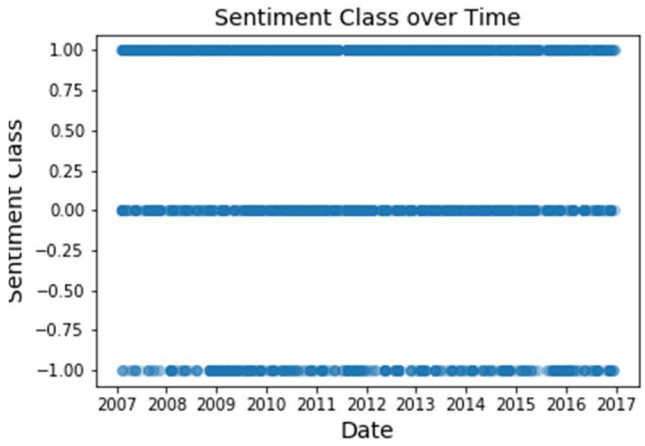
variance in the 10 day market return residual which means if one could predict that metric perfectly then more money could be made from Twitter stock than Agilent Technology stock.

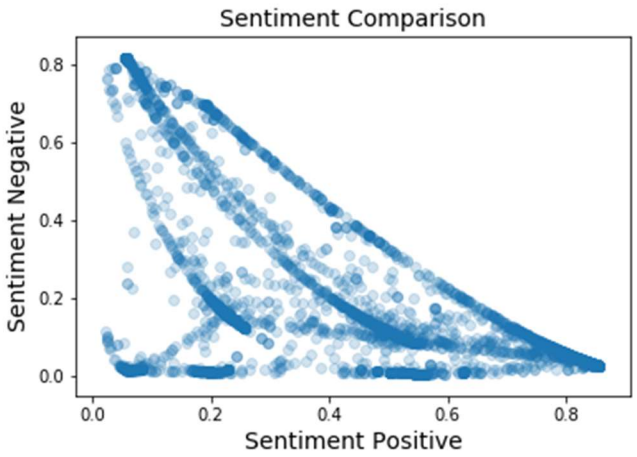


News Sentiment

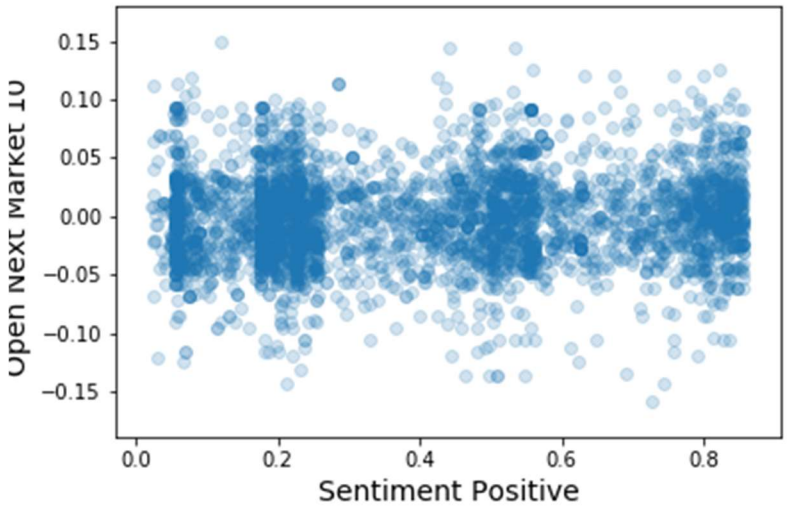
Here the sentiment class is shown over time. The values of 1, 0 and -1 correspond to a news segment with a sentiment of positive, neutral and negative toward Agilent Technology Inc. Notice, that over time there seems to be more positive news segments than negative news segments. Otherwise, the graphic is unremarkable. The sentiment values are explicitly be compared as well. The sentiment positive and sentiment negative values (both of which range from 0 to 1.0) are shown for all news segments related to Agilent Technology Inc. Notice, generally a high sentiment positive values corresponds to a lower sentiment negative value and vice-versa. This makes sense because if an article has a high positive sentiment then one would expect that there is less of a negative sentiment associated with it. However, some news articles are not clearly positive or negative based on these values. Hence, when the sentiment class is chosen there are a fraction of articles that are either neutral or can are not clearly positive or negative toward Agilent Technology Inc. This is a compounding factor that will come into explanations for errors in any predictive model that would be developed later one.

If one were simply trying to find a relationship between news segment sentiment and the 10 day market return residual then the first step would be to visualize the data. Notice, when graphically observing sentiment positive and 10 day market return residual there is no apparent correlation.

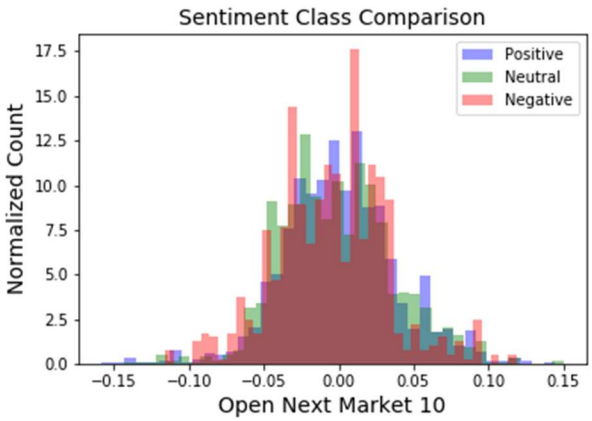


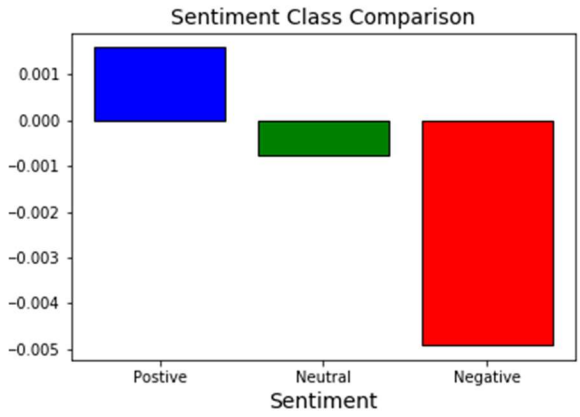


Though it makes sense that these two variables would be correlated this graphic suggests that the correlation is much more complex than a simple linear trend.



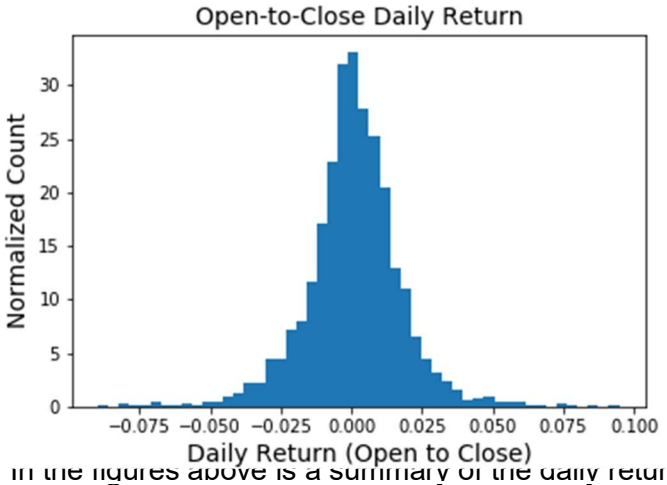
Here the distribution of 10 day market return residuals is graphed for each sentiment class (positive, neutral and negative). Notice, there are slight difference between these distributions however to the naked eye the differences are subtle. This again, speaks to the necessity of more complex modeling to uncover the differences between different types of news reports. The mean of the residual is also calculated for each sentiment class and displayed as a bar chart above. As one logically expects, a positive news sentiment has a positive mean residual and a negative news segment has a mean negative residual. The neutral news segments have a slightly negative residual but not as significantly negative as the negative news segment.

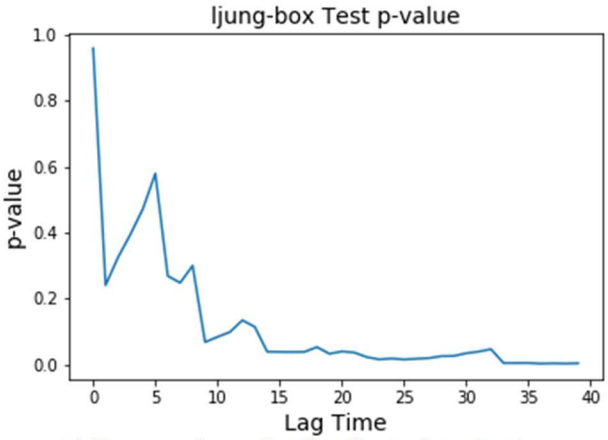




Statistical Analysis

There are a large number of exploratory data analysis techniques to be drawn from the stock market data and the news data. Here a statistical analysis is focused on drawing conclusions between correlations between different features in the data set and reporting on statistical measurements for individual companies. Before exploring these data it would be of interest to understand how short term returns vary day-to-day. Graphing stock market opening or closing prices over time makes data look as if the short term returns or random and independent. This can be checked by performing a ljung-box test and comparing the p-value to a threshold value. If the p-value is lower than the threshold then the null hypothesis fails and therefore the data is not considered random noise.





In the figures above is a summary of the daily returns and the *p-values for the ljung-box test* performed on daily returns for Agilent Technologies Inc. Because of computational limitations performing the test on all companies would have been impractical hence a single company was chosen. Notice, at short lag times <10 the p-value is significant and therefore the null hypothesis fails to be rejected and therefore daily returns at these lag times can be considered random noise. However, at larger time lags the p-value drops which indicates that the null hypothesis will eventually be rejected for large enough lag times. Hence, at long lag times the daily returns do not show random behavior. This trend is statistically interesting and shows a threshold for a timeframe in which daily returns are no longer random and independent.