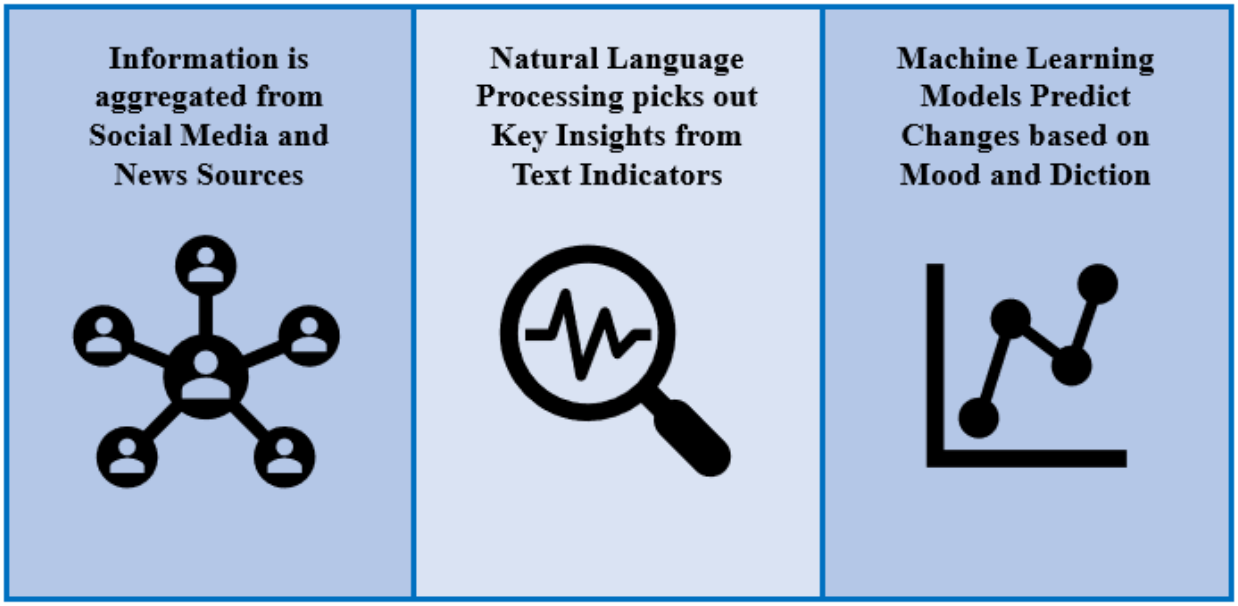
Stock Market Predictions using Natural Language Processing: A Literature Review

Benjamin Anderson

Brigham Young University

Visual Abstract:

Natural Language Processing is used to Predict Fluctuations in the Stock Market



Keywords: Machine learning, Artificial intelligence, Sentiment analysis, Semantic modeling, Twitter, Textual analysis, Computational finance

Stock Market Predictions using Natural Language Processing: A Literature Review

Predicting fluctuations within the stock exchange has been a topic of much study and interest since the New York Stock Exchange was founded in 1871. Everyone with their foot in the stock market from day traders to bankers has been searching for a reliable way to predict these fluctuations so that they can capitalize on the opportunities that they present. Meanwhile, developments in machine learning and artificial intelligence have progressed to create new statistical models that can predict future outcomes with increasing accuracy across many fields and applications. Because even minor changes in the stock market, while trivial to some, can cost corporations and serious investors jobs and millions of dollars, financial analysts have been seeking a way to apply developments in machine learning to help them accomplish their goals regarding the stock market. A method that has caught the attention of financial analysts is called natural language processing, a field of study which seeks to gain insights through diction analysis.

Recent studies boast of success applying natural language processing to the stock market and researchers have investigated the best models and methods for natural language processing. While some comparative studies have been published, most published research focuses on specific areas of natural language processing including semantic models, sentiment analysis, and lexicons. In addition, the studies do not agree regarding which data sources perform the best. This literature review attempts to synthesize these different aspects of natural language processing in the stock market by summarizing the methods that are prevalent in the space today. In addition, we will discuss and analyze the effectiveness of news and social media as data sources. Finally, we will investigate the effective methods of prediction for individual stocks and the stock market in general as provided by recent studies. Regardless of the discontinuity

between methods and data sources, recent studies confirm one fact, natural language processing can be a powerful tool in stock market prediction and consistently increases the accuracy of forecasting models.

Natural Language Processing

In the last decade, methods of natural language processing have grown more sophisticated while the principles remain the same. In natural language processing, analysts and researchers gather, preprocess, clean, and use text to make a prediction. In this section, I will discuss the traditional methods of these principles of natural language processing and the advancements that have been made in the pursuit of more accurate models.

Before any project begins, researchers must be aware of the data that is available. Data has been obtained through various means. Text sources such as news headlines used in a study performed by Khan et al. (2020b) and social media posts from studies such as that performed in Lokesh et al. (2018) are the most common. However, innovative methods such as job message boards have been explored as well in recent years Jordan & Elgazzar (2020). One attribute that makes these sources well suited for prediction and natural language processing is that each of these sources are frequently updated. Daily and even hourly updates are essential to predicting the volatile nature of the stock market. Data from these sources can be obtained through web scraping or by requesting access through an API.

Once the textual data has been collected it must be put into a format conducive to analysis. According to Xing (2018) preprocessing entails three steps: tokenization, stop-word removal, and stemming. Tokenization includes creating a “token” of each work in a data set. For example, the sentence, “Microsoft successfully releases new update.” becomes five separate tokens: “Microsoft”, “successfully”, “releases”, “new”, and “update.” Tokenization allows for

word counts to be made and individual sentiment to be analyzed. The next step in preprocessing is stop-word removal. A stop-word is … well "is". Stop-words are common words that occur frequently in sentences and consequently don't add anything to the analysis. Stop-words increase noise in an analysis and can get in the way of important, albeit less common words. Finally, stemming, is a process through which words with similar meanings are turned into the same word. Words such as "driving" and "drive" are similar in meaning, however, if they are left in their current forms they will be considered as different words. If similar words are left as such then word counts aren't as meaning and techniques such as TF-IDF scores are ineffective. Through stemming researchers and analysts change words such as "driving" and "drive" into "driv" so that the words with similar meanings are counted as the same word. Stemming can easily be accomplished in R or Python.

Preprocessing closely overlaps with data cleaning in many ways. However, it is still important to make sure there are no missing values in the dataset. With numerical data missing data can be filled in using mean or linear imputation if necessary, however, textual data presents challenges to these methods. Therefore, it is important to remove missing values from the dataset so the models will work properly.

The above steps remain consistent for most methods of natural language processing. Most advances in recent years pertain to how researchers preprocess, clean, and analyze text. Natural language processing methods can be categorized as sentiment analysis and semantic models. Sentiment analysis and semantic models deal with either the mood behind the words used or the amount of times a word is used, respectively according to Xing (2018). Sentiment analysis is a common method of natural language processing and is demonstrated in studies such as Atkins et al. (2019), Mishev et al. (2020), & Khan et al. (2020a). In sentiment analysis, each

word is assigned to have a positive, negative, or neutral connotation. The sum of all the connotations in a text show the overall sentiment of the text and can be used to predict whether the stock market or individual stock will rise. One weakness of sentiment analysis is that it relies heavily upon lexicons, which are collections of words that have a sentiment assigned to each word. As discussed by Mishev et al. (2020), there are not many reliable lexicons for financial sentiment analysis. Words such as “bear” and “bull” have a neutral connotation in normal lexicons, when they should be positive or negative in a financial setting. The lexicon that produces the most accurate results was put forth by Loughran and McDonald according to a series of studies performed by Mishev et al. (2020) comparing various lexicons to measure their effectiveness in properly classifying sentiment. While sentiment analysis is a popular method of analyzing text, it has its limitations.

Semantic models differ from sentiment analyses because they rely heavily upon the statistical significance of words in a context and disregard the sentiment behind the words. These are more traditional methods such as bag-of-words and one-hot encoding. TF-IDF scores are one example of a semantic model. TF-IDF stands for “term frequency-inverse document frequency.” Essentially, it is a measure for each word in a dataset that takes the frequency of each word in an observation over the number of times the word is used in the entire dataset. Once numerical values are assigned relationships can be found between certain classifications and words used. With TF-IDF scores words that are common receive a lower score, while rare words that have an impact on classification such as “bear” or “bull” are emphasized in the analysis. Other methods of semantic models include word and sentence encoders as described by Mishev et al. (2020). Word and sentence encoding consider context that cannot be considered in sentiment analysis. For example, the context of the sentence “Microsoft is not successfully releasing new updates"

may be lost in sentiment analysis because the negative sentiment behind “not” is lost in the positive sentiment in “successfully.” With encoding each sentence or each word is given a numerical value giving machine learning models greater insight into the context behind the statements. While these methods are consistent, Ding et al. (2020) has created a new method of encoding which incorporates syntactic and semantic information within the text to analyze event-based information. Previous methods of encoding could understand relationships between words and sentences however, they could also emphasize similarities where there are none. Sentences such as “Steve Jobs has left Microsoft” and “Lauren has left Walmart " would have been categorized as similar events without context, when really, they are not related at all. However, in Ding et al. (2020) they discuss a method of encoding which considers the context of the statement instead of simply the meaning. Ding’s method produces more accurate results than sentiment analysis and previous methods of encoding. Obviously, there are a lot of variables to consider when performing an analysis regarding the stock market using natural language processing, not the least of which is data collection.

Textual Data Sources

Textual data can be pulled from a variety of sources. Within the field of stock market prediction there are two common sources of data: financial news and social media. However, there remains discrepancy regarding which source is best.

Using Financial News

Financial news has been considered a great source of textual data because it does not include a lot of the noise prevalent in social media data and it can easily be accessed through web scraping, a process and where tables and information are pulled from websites. In 2018, two studies were published using financial and world news as their primary data sources. In the study

performed by Atkins et al. (2018), data was pulled from the Reuters US News Archive. An archive meticulously documented to the minute which allowed for real time analysis of the effects of news on the stock market. Atkins et al (2018) suggests that there are two different ways to make predictions regarding changes in the stock market. You can approach changes as a classification problem, in other words, whether the stock increases or decreases in price that day. Or you could approach changes in price as a regression problem and predict the closing price. The results of Atkins and colleagues' study provided evidence that the direction of stock movement upon new information could be predicted with approximately 56% accuracy and the closing price of an asset with approximately 49% accuracy, slightly worse than flipping a coin.

In a similar study performed by Leekha et al. (2018), the problem was broken up a different way. Instead of focusing on classification vs. regression, Leekah et al. (2018) focused on comparing the results of a sentiment analysis and a language processing approach. While Leekah et al. (2018) uses news in much the same way as Atkins et al. (2018), their studies differ in their sources. Atkins et al. use specifically financial news as their primary data source while Leekah et al. (2018) use general news. The difference between these two types of data is that financial news from the Reuters US News Archive relates directly to the stock market while the Top 25 from the Reddit World News Channel used in Leekah et al. (2018) is not focused on financial. These approaches have their own strengths and weaknesses. However, as pointed out in Leekah et al. (2018) investors make decisions off more than just financial news. The general state of the world, including politics and world events, can affect investors tendencies in the stock market as well. The importance of a wide range of news is reinforced by the 85.7% accuracy accomplished in Leekah et al. (2018) when taking a language processing approach. Although the conditions of the studies vary, there is evidence that the market depends more upon

the mood of investors in general and that there is more at play in financial forecasting than just financial news. As a result, studies have confirmed that general world news is better for predicting trends in the stock market as evidenced by Leekah et al. (2018).

Only one study has been performed to see what effects new's headlines have in the stock indexes of specific companies. Results from Katayama & Tsuda (2018) are singular in the field in that they analyze three different hypotheses as opposed to simply testing for greater prediction accuracy. In their studies they can confirm that the stock price of a company rises after positive news comes out regarding it. Secondly, they confirm that articles on the front page have a greater effect on company's stocks than otherwise. Finally, they confirm that the effect of a publication is greater for companies with smaller market capitalization. However, whether it is financial news or general world news both are useful in making predictions regarding the stock market.

Using Social Media

Another popular method of obtaining textual data is through social media. As mentioned above, the quality of social media can be considered lower due to the amount of noise from unrelated statements about the company. However, the mass of information could also be viewed as an advantage because all posts are part of the public sentiment. An advantage to social media is the sheer quantity of data contained within it. Twitter is one of the most popular social media platforms used for obtaining text data. Studies such as those performed by Bollen et al. (2011), Lokesh et al. (2018), and Khan et al. (2020b) are all examples of studies that have been performed using Twitter as a primary source of sentiment data. Because Twitter provides such frequent updates and insights discussed in Khan et al. (2020b), Lokesh et al. (2018) created a mobile application which will inform investors regarding the risk surrounding their investments based on the public's mood.

The studies performed by Bollen et al. (2011) and Khan et al. (2020b) prove that twitter data can be successfully used when making predictions. In fact, Bollen et al. (2011) created a model which can predict with 87.6% accuracy the changes in the Dow Jones Industrial Average. Such a high accuracy is not accomplished without significant cleaning and Bollen et al. (2011) implements a method of data cleaning unique to social media posts that capitalizes on the expressive nature of Twitter. News headlines used in the analyses we discussed earlier seldom use expressions such as “I feel”, “I am feeling”, “I think”, etc. Bollen et al. (2011) capitalizes on this fact by only including tweets that contain such statements. Using this selection method allowed Bollen to home in on the mood of the investors, rather than just the words that they chose. While the results of Khan et al. (2020b) focus more on their success using political news to predict market fluctuations, their findings reiterate the point that using Twitter increased their accuracy by up to 3% consistently.

Stock Market Prediction Approaches

In addition to confirming social media’s reliability as a data source, the combined works of Bollen et al. (2011) and Khan et al (2020b) serve to highlight a juxtaposition in methods of analysis. Bollen et al. (2011) take a broad approach to in analyzing the mood of the population and therefore predict the fluctuations of the stock market in general whereas, Khan et al (2020b) and Lokesh et al. (2018) look at fluctuations in individual stocks. In other words, they look at how sentiment towards a company will affect that company’s stock. The juxtaposition between these approaches is also witnessed in the work of Chiong et al. (2018), Leekha et al. (2018) and Ahuja et al (2015) exemplifying the general stock market prediction approach. While Pagolu et al. (2016), Katayama & Tsuda (2018), Jermann (2017), Khan et al. (2020a) and Jordan & Elgazzar (2020) are examples of individual stock prediction.

Predicting the fluctuations of the stock market in general can be useful when considering investments that span industries or involve diversified portfolios. They can help an investor know if it is a good time to invest in general or if it is a good time to sell. There are many factors at work outside of the stock market which still influence stock prices. For these reasons, predictions regarding the stock market in general rely on news sources more commonly for their predictions. However, the more common of the methods is predicting individual stock price increases and decreases. Predicting individual stock prices is of interest because undiversified portfolios allow for greater returns on investments as well as greater losses. These analyses generally rely more heavily upon social media as a source for their data. Khan et al. (2020a) emphasized this point in their studies proving that social media outperforms financial news when predicting individual stocks. Social media data obtained an accuracy of 80.53% while financial news performed slightly worse with an accuracy of 75.16%.

In addition to data source selection when determining the best way to approach a general or individual stock prediction, the selection of a machine learning model is another important aspect when making predictions regarding the stock market. No consensus has been reached on the topic. With a myriad of machine learning models each with there own strengths and weaknesses along with different measures of performance comparing results is difficult. Regarding predictions of individual stocks Khan et al. (2020a) recommends a Random Forest model because it gives consistent results in all situations. However, the work of Pagolu et al. (2016) produces similar results with a simple logistic regression algorithm. However, regarding predictions made for the stock market in general results from Bollen et al. (2011) suggest that a Self-Organizing Fuzzy Neural Network provided an accuracy of 86.7%. The Fuzzy Neural Network performed only slightly better than the accuracy accomplished using a Multilayer

Perceptron Classifier used in Lokesh et al. (2018). However, it is interesting to note the similarity of the results' accuracies and that both methods were a form of Neural Network.

Conclusion

Predicting the stock market is of great interest to many parties because it represents an opportunity to make a significant profit. However, it presents the investor with equal risk of losing what they have invested. Consequentially, financial analysts are looking to apply machine learning techniques to help them predict stock prices. Natural language processing is at the crossroads of analytics and linguistics and is a promising field for making these predictions. In recent years many studies have applied natural language processing with varying degrees of success to predict fluctuations in the stock market in general and in individual stocks. Natural language processing involves either performing a sentiment analysis or performing statistical analysis on words to determine their significance in each context. While sentiment analysis is a popular method of natural language processing it is often outperformed by semantic methods such as TF-IDF scores.

There are many different sources of textual data that can be used for stock prediction; however, the best sources are those with consistent real-time updates which allow for daily or even hourly predictions to be made. As a result, news outlets and social media are commonly used in natural language processing. The quality and content of data has a large effect on the success of a model or analysis. While these studies such as Khan et al. (2020a) and Pagolu et al. (2016) produce accuracy around 70%, a similar study performed by Jermann (2017) uses specific tweets from company executives to predict the changes in the stock price. The study performed by Jermann yielded little evidence above that of chance, achieving an accuracy of 52.1%. Such a

low accuracy, barely above that of chance, indicates that general sentiment towards a company has a larger effect on the company's stock price than the sentiment of the company's executives.

Finally, making predictions regarding the stock market can be broken up into two subcategories, specific stock predictions and general stock market predictions. The variability and inconsistency between studies makes it difficult to compare because authors test different models, use different lexicons, and utilize different data sources. Generally, for individual stock predictions social media is used because it allows for data be collected on multiple companies because of mass of information available each day. However, for the reasons listed above, results regarding the accuracy of individual stock predictions are incomparable. However, when considering making stock market predictions in general, it is best to use news data which incorporates a wider range of opinions and events needed to predict general stock market trends. The studies of Bollen et al. (2011) and Lokesh et al. (2018) suggest that neural networks perform the best when predicting general stock market trends.

Future research in the field should be focused on comparison of existing methods in a measurable way. The inconsistency between studies makes it impossible to make any more than generalizations regarding which data sources, methods, and models are the most appropriate. In addition, natural language processing should be compared to traditional methods of stock market prediction. In addition, natural language processing could be added to ensemble methods of prediction to see if it improves their accuracy.

References

Ahuja R., Rastogi H., Choudhuri A. and Garg B., "Stock market forecast using sentiment analysis," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 1008-1010.

Abdullah, S. S., Rahaman, M. S., and Rahman, M. S. (2013). Analysis of stock market using text mining and natural language processing. In 2013 International Conference on Informatics, Electronics and Vision (ICIEV), pages 1–6.

Atkins, A., Niranjan, M., and Gerding, E. (2018). Financial news predicts stock market volatility better than close price. The Journal of Finance and Data Science, 4:120–137.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1–8.

Chiong, R., Fan, Z., Hu, Z., Adam, M. T., Lutz, B., and Neumann, D. (2018). A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. Association for Computing Machinery, page 278–279.

Ding, X., Zhang, Y., Liu, T., and Duan, J. (2016). Knowledge-driven event embedding for stock prediction. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2133–2142, Osaka, Japan. The COLING 2016 Organizing Committee.

Frank Z. Xing, Erik Cambria, R. E. W. (2018). Natural language based financial forecasting: a survey. Artificial Intelligence Review, 50:40–73.

Jermann, M. (2017). In Predicting Stock Movement through Executive Tweets.

Jordan, T. and Elgazzar, H. (2020). Stock market prediction using text-based machine learning. In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pages 1–5.

Katayama, D. and Tsuda, K. (2018). A method of measurement of the impact of Japanese news on stock market. Procedia Computer Science, 126:1336–1343.

Khan, W., Ghazanfar, M., Azam, M., Karami, A., Alyoubi, K., and Ahmed, A. (2020a). Stock market prediction using machine learning classifiers and social media, news. Journal of Ambient Intelligence and Humanized Computing.

Khan, W., Malik, U., Ghazanfar, M., Azam, M., Alyoubi, K., and Alfakeeh, A. (2020b). Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Soft Computing, 24(15):11019–11043.

Leekha, A., Wadhwa, A., Jain, N., and Wadhwa, M. (2018). Understanding the impact of news on stock market trends using natural language processing and machine learning algorithms. International Journal of Knowledge Based Computer Systems, 6(2):23–30.

Lokesh, S., Mitta, S., Sethia, S., Kalli, S. R., and Sudhir, M. (2018). Risk analysis and prediction of the stock market using machine learning and NLP. International Journal of Applied engineering Research, 13(22):16036–16041.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., and Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. IEEE Access, 8:131662–131682.

Pagolu, V. S., Reddy, K. N., Panda, G., and Majhi, B. (2016). Sentiment analysis of twitter data for predicting stock market movements. pages 1345–1350.