3-25-2017

Feeling The Stock Market: A Study in the Prediction of Financial Markets Based on News Sentiment

Max Sorto
Georgia Southern University, max_sorto@georgiasouthern.edu

Cheryl Aasheim
Georgia Southern University, caasheim@georgiasouthern.edu

Hayden Wimmer
Georgia Southern University, hwimmer@georgiasouthern.edu

Follow this and additional works at: http://aisel.aisnet.org/sais2017

FEELING THE STOCK MARKET: A STUDY IN THE PREDICTION OF FINANCIAL MARKETS BASED ON NEWS SENTIMENT

Max Sorto
Georgia Southern University
max_sorto@georgiasouthern.edu

Cheryl Aasheim
Georgia Southern University
caasheim@georgiasouthern.edu

Hayden Wimmer
Georgia Southern University
hwimmer@georgiasouthern.edu

ABSTRACT

Researchers are fascinated with predicting the stock market. Even though there is a large amount of supporting evidence that the dynamics of financial markets cannot be predicted, studies that employ creative prediction techniques continue to emerge. This study proposes a sentiment analysis model developed to infer the polarity of news articles related to a company. The process of collecting the dataset, as well as a diagram of the system architecture for the sentiment analysis engine used in this study is provided to readers. Insights from this research and experimental results are used to provide further proof that supports the Efficient Market Hypothesis.

Keywords

Sentiment Analysis, Financial, Markets, Prediction, Text Mining, Opinion, Polarity, Efficient Market Hypothesis

INTRODUCTION

Predicting the stock market is a task that fascinates researchers. It brings together many sciences in the spectrum of statistics to psychology. As data science tools have improved in terms of analyzing and mining information, many predictive applications have followed. It's only natural that researchers keep attempting to predict the stock market, no matter how complex it may seem to model market dynamics. In this study, a sentiment analysis or opinion mining model was developed to infer the polarity of news articles related to a company, and subsequently find a correlation between news sentiment and stock movements. The model proposed in this study is intended to be used by financial investors, hedge fund managers, and other financial experts to better understand the opinions of news article authors. This research also performs a prototype experiment using the proposed system to predict stock market movements. The results of this experiment further prove the efficient market hypothesis.

The dataset used in this research was gathered manually over the course of 30 days. Financial data, such as such as stock price, index price and trading volume, was collected. News articles were collected from the Wall Street Journal as it has a strong voice in the world of finance. Five companies from the Dow Jones Industrial Average (DJIA) were selected for this research. The DJIA was chosen because it has companies that are frequently covered by media. The sentiment analysis tool SenticNet was used to analyze every article and calculate a daily sentiment score for each company.

The proposed sentiment analysis system is explained in detail and a diagram of the system architecture is provided. The results of a prototype experiment using the proposed model for predicting the stock price is carried out. The results of this experiment are congruent with claims observed in the literature that state that the stock market cannot be accurately predicted.

LITERATURE REVIEW

Public opinion plays an important role when it comes to decision-making. Organizations and individuals care about the opinion of others especially when they have stake in a given entity or matter. This research seeks to understand if there is a relationship between the opinions expressed by reputable news reporters and the likelihood of investors to buy or sell stock of a company. The subsequent sections provide a summary of literature related to existing similar research, the state-of-the-art in sentiment analysis, and financial market prediction.

Existing Research

Market prediction based on textual-content sentiment has been attempted before with varying success in results. In Gidófalvi's (2001) experiment, a naïve Bayesian text classifier is trained to predict stock price movement using news articles. The outcome of this experiment implies that the prediction power of this classifier is low, however, there is a strong correlation between the behavior of stock price 20 minutes prior and 20 minutes after news articles become publicly available. According to Gidófalvi, this correlation challenges the idea of a random efficient market. Tetlock's (2007) research finds that there is a strong correlation

between pessimism, or negative opinion, and market movements. The study claims that negative media content "induces a downward pressure" on the price of stocks; this effect has a long lasting impact on the price of small stocks. Tetlock goes on to explain that high and low periods of pessimism also lead to higher volume rates of trading. The results suggest that a predictive model with excess returns of 7.3% per year could be developed, however commission and trading costs would be so high that this operation may not be profitable. In Bollen et al. (2011) prediction of Dow Jones Industrial Average (DJIA) values based on mood is achieved with 87.6% accuracy by analyzing a dataset of 9.8 million tweets. Unlike previous research, their approach looks at more than just polarity (positive or negative) in text; it also looks at mood dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). The authors claim that the general public is as invested in the stock market as financial experts are, and thus their mood has an impact on the value of stocks. Their research finds that the "calmness" in public tweets has the strongest correlation with DJIA movement.

WHAT IS SENTIMENT ANALYSIS?

The popularity of data science and availability of digital text-based content has made *sentiment analysis* flourish as a field of study. *Sentiment analysis* or *opinion mining* seeks to understand public opinion towards an entity. According to Liu (2012), *sentiment analysis* is one of the largest areas of research in Natural Language Processing (NLP) since early 2000 for several reasons. First, there is a wide range of application in many industries (Liu 2012). Second, it poses new research challenges that had not been studied before (Liu, 2012). Third, there's finally a large volume of opinionated data to analyze due to outlets like social media (Liu, 2012).

There is a myriad of real-world applications for sentiment analysis. One of the most common applications and focus for a large amount of studies in sentiment analysis is to automatically detect the polarity of a message - That is, whether the opinion about a topic is positive or negative (Gonçalves et al., 2013). For example, websites that collect user reviews can benefit from automatic sentiment classifiers to increase the accuracy of their reviews. Pang (2008) states that a sentiment analysis tool would correct human error that can occur during the review process when a user accidentally selects a low rating for an otherwise positive review. Another application that seeks to improve human-computer interaction is the use of sentiment analysis to prevent advertisement systems from displaying ads next to content that has a negative tone or has received a lot of negative feedback (Pang 2008, 2012). The work of analysts could be sped up with a sentiment analysis tool that finds opinions of consumers throughout the web and returns a summary of all reviews (Pang, 2008).

New research questions related to sentiment analysis and NLP in general have arisen since the early 2000s. The three levels at which sentiment analysis has been researched at are: document level, sentence level, aspect level. At the document level the goal is to classify an entire document, such as a news article, as either positive or negative (Pang et al., 2002). Similarly, at the sentence level the goal to classify a sentence as positive, negative or neutral (Liu, 2012). Aspect level techniques have a more granular and precise goal: to identify the sentiment as well as the target of opinion which is a lot more difficult than document-level and sentence-level analysis (Liu, 2012). So for example a two-part sentence such as "Although the movie was bad, he is still my favorite actor" is positive about the actor but negative about the movie. Additionally, Jindal and Liu (2006) state that there are two types of opinions: regular opinions and comparative opinions. A regular opinion contains the sentiment of a single topic or entity. A comparative opinion expresses sentiment about multiple entities based on a common trait. For example, "The PS4 is a good gaming system" is a regular opinion, and "The Xbox One performs better than the PS4" is a comparative opinion about the two console's performance with preference for the first console.

Due in part to the growth of the World Wide Web, there is now a large volume of data available to work with. Pang (2002) states that machine learning algorithms have been improved thanks to the increase in availability of datasets. Another major contributor to data availability is social media. Individuals and organizations rely on the content of social media posts, blogs, tweets, comments, etc. for decision making - reducing the need for traditional surveys, polls and focus groups (Liu, 2012). That being said, organization have also found great value from performing sentiment analysis on internal data collected from emails and surveys (Liu, 2012).

What tools are available for Sentiment Analysis?

Two of the main techniques for performing sentiment analysis are machine-learning approaches and lexicon-based approaches (Gonçalves et al., 2013). The machine learning approach involves training with a set of data that has a given output - this output is then used to generate predictions for any new input data. Lexicon-based approaches use dictionaries of words annotated with the word's semantic orientation to determine polarity. Over the next few paragraphs a summary of tools suggested by Gonçalves et. al (2013) as state-of-the art sentiment analysis methods available.

LIWC

Linguistic Inquiry Word Count (LIWC) is a commercial tool that counts words and sorts them into psychologically meaningful categories (Tausczik, 2009). LIWC, like the basis of most sentiment analysis tools, involves comparing words to a dictionary of words that belong to categories. Tauszcik (2009) explains that emotion word categories in dictionaries were selected by human judges; the process of sorting words into categories has been a central piece to LIWC since the 1990s (Tausczik, 2009).

SentiStrength

Gonçalves et al. consider SentiStrength to be the state-of-the-art in social network sentiment analysis. They go on to explain that this method relies on LIWC's dictionary as well as other features that are specific to the context of online social networks. Words are assigned both a positive and negative score as a result of a psychological study that states that human can experience both positive and negative emotions simultaneously (Norman et al., 2011). SentiStrength uses a lexicon-based approach that works well on many social web texts without requiring any training data (Thelwall, 2013). This tool is freely available for educational and research purposes in a Windows and Java version; however, commercial users are required to pay for the Java version.

SentiWordNet

SentWordNet is a free-to-use tool that is popular in opinion mining research (Baccianella et al., 2010). The tool uses a dictionary called WordNet, which groups words into sets of synonyms called synsets (Gonçalves et al., 2013). SentiWordNet gives three scores to each synset: a score for positive, a score for negative, and a score for objective (or neutral). The three scores are in the range of [0-1] and add up to 1. For example, a synset of the word "*terrific*" would score 0.850 for positivity, 0.0 for negativity, and 0.150 for objectivity. Non-zero values for all three categories are also acceptable, as it indicates that a synset has all three types of opinion to a certain degree (Cambria et. al 2010). As of this writing, the latest release is SentiWordNet 3.0 and it is freely available for non-profit research (Baccianella et al, 2010).

SenticNet

SenticNet is a public sentiment analysis and opinion mining tool that takes advantage of artificial intelligence and semantic web techniques (Cambria et. al 2010). The development of SenticNet was inspired by SentiWordNet, which provides opinion polarity at a syntactical level; however, SentiWordNet cannot understand common sense expressions such as "on cloud nine" (Cambria et. al 2010). Cambria et. al explain that SenticNet relies on the Open Mind Common Sense project, which has been collecting common sense knowledge since 2000 to provide intuition to artificial intelligence environments. Unlike SentiWordNet's approach of assigning scores to all three of the opinion-related properties of a synset, SenticNet assigns only one score for each concept based on its polarity in the range of [-1,1] (Cambria et al., 2010). As of this research, the latest version of this tool is SenticNet 4.0 and it is available for download at no cost from the developer's website.

Emoticons and Emojis

Detecting sentiment based on emoticons is one of the quickest and easiest ways to detect the polarity of a message (Gonçalves et al., 2013). An emoticon is a text representation of a facial expression that conveys an emotion. Emoticons are most commonly found informal texts, naturally. Emojis are more commonly used than emoticons nowadays as they have been integrated to most digital keyboards and operating systems ever since the Apple iPhone first included them in 2010 (Novak et. al 2015). In comparison, emoticons are centered on the emotions of the writer, whereas emojis are based around ideas. Depending on the dataset being analyzed, emoticons and emojis might not indicate much about a text, as it would be in the case of a news article which contains more formal writing.

What is Market Prediction?

Many organizations and academic institutions are interested in predicting the stock market. Approaches to predicting the stock market are either technical or fundamental. Technical analysis is based on historical financial data, whereas fundamental analysis is based on timely corresponding data such as organizational changes and product releases. More fundamental analysis surfaced as financial textual data (news articles) became readily available on the web (Gidófalvi, 2009). Malkiel et al. (2003) suggest that once any predictable pattern is made publicly available there is a rush to exploit it, which makes the pattern disappear.

Efficient Market Hypothesis

The efficient market hypothesis (EMH) was developed by Eugene Fama, who explains that stock prices are traded at their fair price, which is a representation of past and new information regarding that company. The EMH affirms that stock price cannot be predicted with more than 50% accuracy (Qian et. al, 2007). Qian et. al explain that the volatility of stock markets is explained by EMH as a "random walk", meaning that every day prices reflect what's going on that day; Malkiel et al. (2003) suggests

that the odds of making return on investment are the same if stocks are picked at random or by an expert; their research also suggests that once a predictability patterns disappear once findings are published. Even though the EMH theory says prediction attempts are futile, the lucrativeness of solving this problem is enough of a motivation for research to continue. Towards the beginning of the twenty-first century, critics of the EMH theory began to argue that the movement of stocks was partially predictable based on historical patterns and (Malkiel et al., 2003). Additionally, Malkiel et al. (2003) believe that most studies that show predictable patterns are simply the result of using the right sample with the right technique to produce statistically significant results.

Short-Term Momentum

Malkiel et al. (2003) state that as behavioral psychology became more influential, momentum gained more interest as an explanation of stock movement, as opposed to the "random walk" views of EMH. Financial experts argue that a stock moving in the same direction for a successive period will impact the predictability of future behavior (Malkiel et. al, 2003). This means that when a stock price continuously increases there is collective impulse for investors to buy that stock, which in turn continues to make the price rise; Psychologists and economists describe this impulse as a "bandwagon effect" (Malkiel et. al, 2003). An example of this bandwagon effect can be observed during the late dotcom boom in the late 1990s. During this time companies would even add "dotcom" to their name simply to raise investor interest (Shiller, 2000). Malkiel et al. reaffirm that even though these short-term momentum strategies had good returns on investments in the 1990s, they did not hold well in the 2000s, thus not being to prove the market as inefficient.

Seasonal Anomalies

Research suggests that certain times of the year, month, and week have predictable outcomes for the market. For example, French (1980) explains that the calendar time hypothesis expects Monday returns to be three-times higher than returns any other day of the week. Lakonishok and Smidt (1998) report that there are also patterns on returns towards the end of month and around major holidays. The most commonly known seasonal anomaly is explained by Haugen and Lakonishok in their book titled The Incredible January Effect, where they overview the unusually high stock market returns in the month of January. Haugen and Lakonishok explain that smaller stocks provide a greater return on investment than large stocks when purchased at end-of-year and sold in the first ten trading days of January. Haugen and Jorion (1996) published research work a decade after the publication of The Incredible January Effect that proved this effect was still going strong. On the other hand, Malkiel et al. (2003) suggests that the January effect is bound to wear off as investors will be competing to buy and sell stock earlier and earlier to key days, removing the possibility of excessive returns.

Valuation Parameters

People look to invest in the market with the hopes of earning profit. That is why information such as dividend to stock price (dividend yields) and price-earnings ratios have been analyzed in hopes of finding patterns that indicate who to invest in. Fama and French (1998) performed studies on the predictability of stock movements based on dividend yields. Their research finds that initial dividend yield of the market index can predict future earnings with up to 40 percent accuracy. However, Malkiel et al. (2003) mention that corporations have changed their dividend behavior over time, such as providing share repurchase programs, has made this method of prediction inefficient. Another valuation parameter that was studied by Campbell and Shiller (1998) is price-earnings ratios. Coincidentally, their research also suggests that price-earnings ratios also predict future earnings with up to 40 percent accuracy (Campbell and Shiller, 1998). That being said, it's worth considering that investors have observed that this pattern has not held in other time periods, where the opposite of predicted earnings occurs (Malkiel et al., 2003).

METHODOLOGY

The subsequent subsections highlight all aspects of the work performed in the development of a proposed model for predicting movements of the stock market based on the sentiment of news article authors. The process for collecting the dataset used in this research is explained. The model has a sentiment analysis component that determines the polarity (positive or negative) of the content of news article. The proposed model is intended to be used by financial investors, hedge fund managers, and other financial experts in the development of better investment strategies that account for the opinion of reputable news authors. Lastly, a prototype experiment using the proposed model and collected dataset is carried out to find a correlation between stock movements and news sentiment.

Collecting the Dataset

Gathering the data for this research was a month-long process. A set of five companies traded in the Dow Jones Industrial Average (DJIA) was chosen for this study. News articles about these companies was collected for a period of 30 days. During

this same period, market data for these companies was also gathered. The dataset for this study was manually gathered; however, based on subsequent experience, the news articles can be automatically collected using screen-scraping techniques.

Selecting Stocks

The DJIA companies selected for this study were American Express, IBM, Intel, Microsoft, and Nike. These five companies were the top gainers in the index the day we began collecting data. The reasoning behind using the top gainers for this study was the assumption that these stocks are popular and thus there will be plenty of new content available to analyze. Furthermore, it was agreed that this choice was justifiable because companies selected did not belong to the same industry. Table 1 shows the companies picked along with their stock tickers. The market data collected each company every day was closing price, index price, and trading volume. Market data was collected from the NASDAQ stock market official website. The collecting of the index price and trading volume was with the purpose of finding additional correlations between news sentiment and market behavior that challenge the efficiency of markets.

Company	Stock Ticker
Microsoft Corp.	**MSFT**
Intelligent Business Machines Corp.	**IBM**
Nike Inc.	**NKE**
Intel Corp.	**INTC**
American Express Co.	**AXP**

Table 1: Top DJIA Gainer Stocks on September 27, 2016

Collecting News Articles

A single news source was used for analysis due to the manual nature of our collection process; however, the decision-making approach produced in this study is intended to work with any source of news articles. The news source picked for this article was The Wall Street Journal because it is a well-established and business-focused source. Every day a maximum of five articles tagged with the stock ticker for each company was collected as text files in corresponding folders for each day. The text files only contained the headline of the news article as the first line followed by the article's text contents.

Data Preparation

Market data such as closing price, index price, and trading volume were collected in a spreadsheet. To get additional insight into the behavior of the market, each day also included values for the previous day. Three more variables were calculated based on data from the current and prior day which are price movement, trading volume movement, and movement of the DJIA; the results for these were labeled as UP or DOWN. The last variable added to the dataset was the daily sentiment score which was calculated by running all articles corresponding to a company and day through a sentiment analysis system. More details on how the sentiment analysis system works is provided in the **System Architecture** subsection. Our aim was to predict stock movement and trading volume movement; thus, these were each used as dependent variables for testing the prediction accuracy of our system.

System Architecture

To calculate the daily sentiment score for each company, all articles were processed through a Python-based sentiment analysis extraction system. The system is composed of two main parts: a summarization tool and the SenticNet 4.0 API. News articles are reduced to topic sentences, which are then split into individual words that are passed to the SenticNet API. The reason a summarization of the articles is used as opposed to the entire article is to remove excess words that can introduce noise. The summarization tool retains the main message of the articles without excess text. SenticNet assigns a polarity score to each word and then the average score of all words is calculated for the news article. The average of all news articles in a day becomes the daily sentiment score.

Summarization Algorithm

The summarization algorithm used in this study is a variation of the one created by Shlomi Babluki (https://thetokenizer.com/2013/04/28/build-your-own-summary-tool/). The user is asked to enter a stock ticker and a date which indicates to the tool what news article *.txt* files to look for and pass to the summarization algorithm. The entire article is parsed as a string, which is then split into sentences and the intersection value of these sentences is stored in a matrix. This matrix then holds the intersection score for each sentence, which can then be used to determine which sentences are most descriptive. These descriptive sentences are then stored in a key-value dictionary, where the sentence is the key and the value

is the score calculated (Babluki, 2013). Once each paragraph's most descriptive sentence is identified, the summarization algorithm puts them in a logical order to build a summary of the article.

SenticNet 4.0

The next step performed by the system is passing the summary provided by the summarization algorithm for each news article to SenticNet. The reason SenticNet was chosen over other tools that do sentiment analysis was due to the availability of this package through the Python Software Foundation as opposed to individual user accounts on GitHub for SentiStrength and the authors' familiarity with Python To do this, the summary of the article is split into a list of words. Then, using a for-loop, each word is converted to lowercase for better compatibility with the dictionary in SenticNet. The SenticNet dictionary contains a score for each concept associated with each synset based on emotional label. In that same for-loop, each word is passed to SenticNet to calculate a polarity score which is in the range of [-1, 1] where -1 is very negative emotion and 1 very positive emotion. If the word is not found in the SenticNet 4.0 dictionary, then that iteration of the loop gets skipped. For every successful iteration, a counter variable is increased by one and the polarity score of the word is added to a running total polarity score sum. The total polarity score sum then gets divided by the value of the counter which produces the sentiment score for the article. This process is done for each of the articles available in a single day; the average sentiment score of all articles is accepted to be the daily average sentiment score for a company. The entire process is summarized in Figure 1.
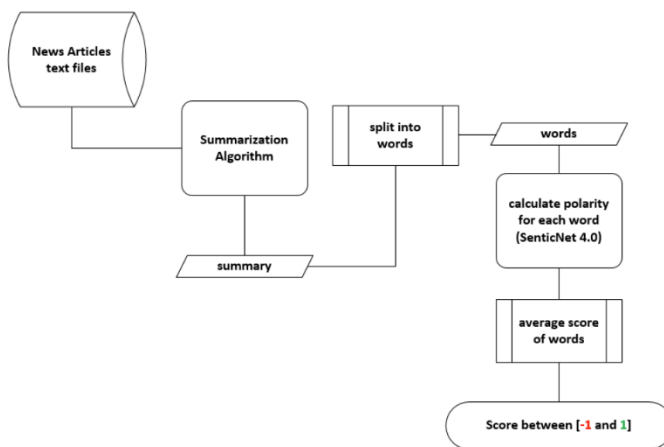

Figure 1: System Architecture of News Sentiment Extraction Tool using Summarization Algorithm.
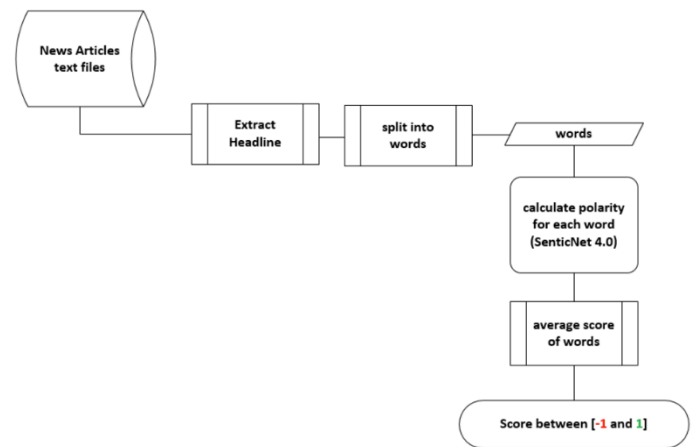
Figure 2: System Architecture of News Sentiment Extraction Tool using only Headlines.

The sentiment analysis system was modified to also calculate the sentiment score using only the headlines of news articles. The summarization algorithm is not included in this modified version as it is not necessary. This alternative approach was based on a study that suggests that headlines are more straight-to-the-point, as opposed to analyzing the entire contents of the article which could cause more noise and affect the results provided by most sentiment analysis tools (Nassirtoussi et al., 2014). The system architecture of the modified system is illustrated in Figure 2.

ANALYSIS

We hypothesize that the opinion of news authors influences the movement of stocks and the trading volume. A prototype experiment was performed using the data collected to test the prediction capabilities of the sentiment analysis system. This experiment seeks to find the percentage of correctly classified days. A correct classification is one in which the sentiment score for that day (UP or DOWN) matches financial data movements. Additionally, the experiment seeks to challenge the Efficient Market Hypothesis. As expected, the results show that markets are indeed unpredictable, maintaining the EMH.

Two test runs were performed; in the first run the sentiment score was calculated using the news article summaries, and in the second run the sentiment score was calculated using only the headlines. For each test, the prediction accuracy of the system was calculated by running a logistic regression test comparing the daily sentiment scores to the movement of the stocks and the movements of trading volume. The logistic regression was chosen because our variables are categorical in nature, they are either UP or DOWN, and as such a logistic regression can be used to predict the likelihood of an outcome based on the input variables (Dietrich et al., 2015). In the subsequent subsection, this process is described in further detail.

Prototype Experiment

There were days during the data collection period were *The Wall Street Journal* did not publish articles for some of the companies observed. To get around this issue, if there were no articles that day, then the most recent current score was assigned until articles were published again. After this, the days when the market is closed were dumped from the results as they do not provide a movement in stock price. A total of 115 instances were analyzed for both runs.

Next, with a working news article sentiment analysis tool, the sentiment score was calculated for each day. It was observed that in the scale of emotions, the scores were mostly on the positive end of the spectrum. It was decided that if a score was above 0.2 it was positive, otherwise negative. This cutoff score was decided because it maximized the total amount of correctly classified instances.

Sentiment Analysis System	Predicting	Correctly Classified Days	ROC
Summaries	Stock Movement	**75%**	0.718
Summaries	Trade Volume Movement	**56%**	0.499
Headlines	Stock Movement	**75%**	0.735
Headlines	Trade Volume Movement	**56%**	0.503

Table 2: Percentages of correctly classified days using SA summarization tool and headlines.

After calculating the sentiment scores and movements of stock and trade volume, the dataset was loaded to an R-language tool. A logistic regression test was used to calculate the prediction accuracy of the sentiment analysis system. The results for each test run, one for summaries and one for headlines, are presented in the Table 2.

In the run using article summaries, the tool shows 75% prediction accuracy for stock movement based on sentiment score and our additional independent variables. The prediction accuracy for trade volume movement was 56%. In the run using only headlines the accuracy was curiously also 75% and 56% for stock movement and trade volume respectively. The Receiver Operating Characteristic (ROC) was slightly higher when using the headlines only run. Future studies expanding on this project should test the system with a significantly larger dataset to see if the similarity in prediction accuracy between both variations of the system remain. The percentages for our test runs do not show strong prediction capabilities; the results are weak, which further support the Efficient Market Hypothesis.

CONCLUSION AND FUTURE WORK

This study proposed a sentiment analysis system developed to infer the polarity of news articles related to a company with the purpose of predicting the stock market. The dataset included news articles from the Wall Street Journal and financial market data from the NASDAQ. The sentiment analysis system used a summarization algorithm and the SenticNet 4.0 API. A modified version of the system was also tested using only news headlines, which omitted the summarization portion of the original system. Both versions of the system were used to carry out a prototype experiment, for which the results further support about the unpredictability of markets, maintaining EMH. The source code for the system can be found on GitHub at https://github.com/maxsorto/SentiTool/blob/master/sentitool.py.Future directions includes testing the system using a significantly larger dataset to better determine the prediction accuracy and that we should include stocks that were not necessarily the biggest gainers. The dataset could include more news article sources to avoid having days in the dataset with no sentiment score. Additionally, the sentiment analysis engine could benefit from further testing and the development of revisions which provide more accurate sentiment scores. Finally, a comparison of the architecture on a news article summary versus using the whole text of the article to determine whether the summarization algorithm is a valuable part of the system architecture can be made.

REFERENCES

1. Babluki, S. (2013). Build your own summary tool! Retrieved December 01, 2016, from https://thetokenizer.com/2013/04/28/build-your-own-summary-tool/
2. Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In LREC, vol. 10, pp. 2200-2204. (2010).
3. Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2, no. 1 (2011): 1-8.

4.   Cambria, Erik, Robert Speer, Catherine Havasi, and Amir Hussain. "SenticNet: A Publicly Available Semantic Resource for Opinion Mining." In AAAI fall symposium: commonsense knowledge, vol. 10, no. 0. 2010.
5.   Campbell, John Y., and Robert J. Shiller. "Valuation ratios and the long-run stock market outlook." The Journal of Portfolio Management 24, no. 2 (1998): 11-26.
6.   Dietrich, D., Heller, B., & Yang, B. (2015). Data science & big data analytics: Discovering, analyzing, visualizing and presenting data. Indianapolis, IN: Wiley.
7.   Dodds, Peter Sheridan, and Christopher M. Danforth. "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents." Journal of Happiness Studies 11, no. 4 (2010): 441-456.
8.   Fama, Eugene F., and Kenneth R. French. "Dividend yields and expected stock returns." Journal of financial economics 22, no. 1 (1988): 3-25.
9.   French, Kenneth R. "Stock returns and the weekend effect." Journal of financial economics 8, no. 1 (1980): 55-69.
10. Gidofalvi, Gyozo, and Charles Elkan. "Using news articles to predict stock price movements." Department of Computer Science and Engineering, University of California, San Diego (2001).
11. Gonçalves, Pollyanna, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. "Comparing and Combining Sentiment Analysis Methods." Proceedings of the First ACM Conference on Online Social Networks - COSN '13, 2013. doi:10.1145/2512938.2512951.
12. Haugen, Robert A., and Josef Lakonishok. The incredible January effect: The stock market's unsolved mystery. Business One Irwin, 1988.
13. Haugen, Robert A., and Philippe Jorion. "The January effect: Still there after all these years." Financial Analysts Journal 52, no. 1 (1996): 27-31.
14. Jindal, Nitin and Bing Liu. Mining comparative sentences and relations. In Proceedings of National Conf. on Artificial Intelligence (AAAI-2006). 2006.
15. Lakonishok, Josef, and Seymour Smidt. "Are seasonal anomalies real? A ninety-year perspective." Review of Financial Studies 1, no. 4 (1988): 403-425.
16. Liu, Bing. Sentiment Analysis and Opinion Mining. Morgan & Claypool, 2012.
17. Malkiel, Burton G. "The efficient market hypothesis and its critics." The Journal of Economic Perspectives 17, no. 1 (2003): 59-82.
18. Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. Expert Systems with Applications, 41(16), 7653-7670.
19. Norman, Greg J., Catherine J. Norris, Jackie Gollan, Tiffany A. Ito, Louise C. Hawkley, Jeff T. Larsen, John T. Cacioppo, and Gary G. Berntson. "Current emotion research in psychophysiology: The neurobiology of evaluative bivalence." Emotion Review 3, no. 3 (2011): 349-359.
20. Novak, Petra Kralj, Jasmina Smailović, Borut Sluban, and Igor Mozetič. "Sentiment of Emojis." Plos One 10, no. 12 (2015). doi: 10.1371/journal.pone.0144296.
21. Pang, Bo, and Lillian Lee. Opinion Mining and Sentiment Analysis. Hanover, MA: Now Publishers, 2008.
22. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). 2002.
23. Qian, Bo, and Khaled Rasheed. 2007. Stock market prediction with multiple classifiers. Applied Intelligence 26, (1) (02): 25-33
24. Shiller, Robert J. Irrational Exuberance Ed. 2. Princeton University Press, 2005.
25. Tausczik, Y. R., and J. W. Pennebaker. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." Journal of Language and Social Psychology 29, no. 1 (2009): 24-54. doi:10.1177/0261927x09351676.
26. Tetlock, Paul C. "Giving content to investor sentiment: The role of media in the stock market." The Journal of Finance 62, no. 3 (2007): 1139-1168.
27. Thelwall, Mike. "Heart and soul: Sentiment strength detection in the social web with sentistrength." Proceedings of the CyberEmotions (2013): 1-14.
28. Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." In Proceedings of the ACL 2012 System Demonstrations, pp. 115-120. Association for Computational Linguistics, 2012.