MPC MAJOR RESEARCH PROJECT

HIDING IN PLAIN SIGHT:
SENTIMENT ANALYSIS AND THE EFFICIENT MARKET HYPOTHESIS

DAVID LITWIN

Supervisor: Dr. Charles H. Davis

The Major Research Paper is submitted

in partial fulfillment of the requirements for the degree of Master of Professional Communication

Ryerson University Toronto, Ontario, Canada

August 19, 2019

Author's Declaration

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PAPER

I hereby declare that I am the sole author of this Major Research Paper and the accompanying Research Poster. This is a true copy of the MRP and the research poster, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this major research paper and/or poster to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP and/or poster by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP and/or my MRP research poster may be made electronically available to the public.

Abstract

The stock market is a notoriously complex and unpredictable system, and because of this has always been an alluring subject for academic research seeking to make the unpredictable more predictable. This major research project is no different as it aims to quantify the predictive value of financial sentiment, determine which sentiments are most meaningful, when they are most meaningful, and if meaningful sentiment varies depending on type of stock. To pursue these goals, the project finds its theoretical footing in Eugene Fama's Efficient Market Hypothesis and Daniel Kahneman's Prospect Theory. However, the methodological component of this project enters into emerging territory as it employs sentiment analysis and machine learning, which have only recently been made possible by advances in technology and communications practices. Specifically, through the use of the Loughran-McDonald dictionary for financial sentiment, corporate press releases were analyzed and tested using a Random Forest machine learning model. The results from this project show that financial senitiment found in press releases does provide a slight predictive edge, however the sentiments responsible for that edge vary based on type of stock, type of fluctuation being predicted, and timeframe.

Acknowledgements

Looking back on this major research project, I still don't know how I decided on sentiment analysis, machine learning, and stock market predictions as my topic. However, what I do know is that completing this project would not have been possible without the help of so many wonderful and inspirational people.

I would first like to thank all of the faculty at Ryerson who made pursuing this degree such a terrific experience. Dr. Charles Davis, as my supervisor, I want to thank you for all of the inspirational meetings, your unwavering support, encouragement, and so much more. Combined, all of this allowed me to discover and explore a topic I am truly passionate about, and for that I am most thankful. To my second reader, Dr. Robert Clapperton, thank you for catching all the lingering typos, teaching us the foundations of research methods, and allowing me to get some real-life experience in natural language processing. And to our "GPD guy" Dr. Matthew Tiessen, thank you for always finding the time to meet with all of us in the program, you care and kindness throughout this year has helped us all in reaching the finish line.

Finally, a big thank you to my family, friends, and the MPC class of 2018/2019, who really are both family and friends at this point. I want to thank you all for always listening to my ramblings, offering your insight, and being there for moral support. Because of the people mentioned here and so many more, this project was a joy to complete and is something I will cherish for a long time.

Table of Contents

List of Tables

List of Figures

List of Appendicies

Introduction

In 2013, Eugene Fama became a Nobel Laureate in Economics for his body of work regarding the Efficient Market Hypothesis (EMH) and the empirical evidence for determining asset pricing. However, in 2002 the same prize was awarded to Daniel Kahneman, whose work on the integration of psychological modelling and economics runs in opposition to Fama's notion of an efficient market. Despite the incongruity of their findings, there is common ground in the aspirations of both Fama and Kahneman to make sense of how information influences the stock market. For Fama, the EMH interprets market fluctuations through a financial lens, while Kahneman's Prospect Theory (PT) provides insight through the use of psychological factors.

The following literature review will first outline Fama's EMH and its implications. Then the literature review will tackle the conflicting behavioural forces that influence both investors and the market through Kahneman's PT. Finally, at the juncture between Fama and Kahneman lies the use of sentiment analysis, which combines financial and psychological factors. As a result, sentiment analysis provides a waypoint into understanding the relationship between the interpretation of market news, and what effect the reporting of said news has on the fluctuation of stock prices.

The purpose of this study is, therefore, to quantify the abstract psychological factors/sentiments present in corporate press releases in order to test them for their ability to predict financial outcomes. It stands to reason that if a press release reports on important corporate events, the sentiment surrounding those events should translate into a timely and measurable response by the company's stock price. Despite this logic, stock prices are notoriously unpredictable. Numerous factors influence stock price fluctuations, and therefore one cannot expect there to be one universal explanation that explains stock prices in their entirety. In addition to quantifying psychological factors, and testing for predictive accuracy, this study also

aims to determine which sentiments contribute most to a stock price's fluctuation, and whether or not those sentiments differ between types of stocks.

While the study of financial markets to uncover explanations that shed light on how and why stock prices change is not a new endeavour, the use of sentiment analysis to do so is. Only recently has technology progressed to a point where collecting and analyzing large quantities of text has become feasible. Methods used in this study such as web scraping and machine learning have only come into existence, and subsequently, prominence relatively recently. Likewise, the proliferation of information published online and made openly accessible has increased dramatically in recent years as well. The coalescence of these technological advancements has opened the door for an emerging area of study.

Literature Review

The framework and direction of this project relate back to the aforementioned conflicting theories of Fama and Kahneman. In its most distilled form, the source of conflict entails whether or not the identification of influential information over the stock market is actionable. Fama's EMH would argue that this is not the case since publicly available information is immediately factored into the price of a stock. Meanwhile, Kahneman's PT would argue that the characteristics of human decision making and the psychological influences at play allow for the possibility of acting on new information.

It should also be noted that the use of Kahneman's PT is not to disprove Fama's EMH. On the contrary, Fama's EMH must be true to some degree otherwise new information would have no impact on the stock market, and sentiment analysis of press releases would be fruitless. Therefore, this literature review will first discuss the core tenets of the EMH and then discuss Fama's categories of information, which allow for the existence of information that may not be as efficiently incorporated into market prices. This literature review will then discuss Kahneman's PT and use it to critique the EMH and offer a psychological explanation as to why the market may be more inefficient at processing certain types of information. Specifically, the discussion of PT will highlight the subjectivity of investor decisions and discrepancies in their interpretation of information.

Last, this literature review will discuss the more practical side of this study, which involves the use of sentiment analysis and machine learning. Discussion of sentiment analysis will draw briefly on its historical context, the types of information that have been interpreted using sentiment analysis, the various methods employed to improve its predictive capabilities and some of the conclusions that have been drawn. Finally, a review of machine learning and the Random Forest model used for categorical prediction will be covered.

The Efficient Market Hypothesis

The essence of this theory states that the price of tradable assets listed on the stock market consistently reflects all public information (Fama, 1970; 1991). The implication here is that the average investor cannot consistently and knowingly turn a profit in the market because an efficient market excludes the possibility of an informed advantage. As Timmermann and Granger put it, "in its crudest form… the returns from speculative assets are unforecastable" (2004, p. 15). It should also be noted that Timmermann and Granger's use of the word "speculative" refers to the difficulty of valuing an asset in a strictly financial context. Later in this literature review speculative assets will be touched on again, but within the context of using investor sentiment for forecasting purposes. For now however, Timmermann and Granger find that in a financial context, any positive return on investment is, therefore, more a product of unactionable factors, which causes the market to go up as a whole thus netting the investor a gain solely based on the time that has elapsed since their initial investment.

According to Fama, an efficient market depends on the mitigation of financial barriers to trading or acquiring information, and agreement among investors on the interpretation of the information (2004; 1991). While these factors are important for the efficiency of the market there exists some flexibility. Fama notes that "disagreement among investors about the implications of given information does not in itself imply market inefficiency unless there are investors who can consistently make better evaluations of available information" (1970, p. 388). The best example given in relation to this involves a longitudinal study of the performance of professional mutual fund managers when compared against the performance of the index over a ten-year period. This seminal study conducted by Michael Jensen found that the vast majority of professionally

managed funds underperformed (as cited in Fama 1970; 1991) and that investors were better off passively investing in the market as a whole.

To better illustrate this point, Fama also discusses the three categories of information. These are the "weak form," which involves predictions based on past returns, "semi-strong form," which are predictions based on all public information, and "strong form," which is similar to semi-strong, but with the addition of private information (Fama, 1970; 1991). Of these categories, weak form and semi-strong form appear to support the EMH, while the strong form category serves "as a benchmark against which deviations from market efficiency (interpreted in its strictest sense) can be judged" (Fama, 1970, p. 415). Specifically, the strong-form tackles instances of "monopolistic access" to information, which result in asset price anomalies (Fama, 1970, 1991). It is still unclear as to whether or not professional investors do in fact have "monopolistic access" to private information (Fama, 1991). If professional investors do have access to private information, then one would assume that they would be able to consistently and knowingly beat the stock market. This however, contradicts the empirical evidence presented by Jensen, which shows that this is not the case.

Going forward, there have been numerous attempts to uncover market inefficiencies and contradictions to the EMH with varying levels of success. The most notable of which is the "January Effect," which found "the existence of seasonality in monthly rates of return on the New York Stock Exchange" (Rozeff & Kinney as cited in Rossi, 2015, p. 288). However, "there is no single, unified point of view on the relationship of the EMH to calendar effects" (Rossi, 2015, p. 293). This "fragmentation" of findings that Rossi describes is due in part to the fact that new predictive theories are self-destructive (Timmermann & Granger, 2004). Namely, once a theory becomes public knowledge the EMH acts and reflects that theory in its corrected price.

Along the same vein, Fama describes it such that, "if a past anomaly does not appear in future data, it might be a market inefficiency, erased with the knowledge of its existence" (Fama, 1991, p. 1593). Therefore, even if an inefficiency were to be found, once made public it would inevitably be adopted and turned into an efficiency.

An Inefficient Market Hypothesis

On the other hand, an inefficient market hypothesis would assert that the market does not reflect all public information in asset prices, at least not right away. In an inefficient market, it would be theoretically possible for an investor to make informed and successful trading decisions that could consistently outperform the index. Conversely, if the market were perfectly efficient "there would be no incentive for professionals to uncover the information that gets so quickly reflected in market prices" (Malkiel, 2003, p. 80), nor would there be any incentive to actively trade. In essence, an inefficient market means prices are not always perfect, leading to assets being dramatically overvalued or undervalued, as it was during the "crash of 1987" or the "Internet bubble" (Malkiel, 2003). While the "true value will win out in the end" (Malkiel, 2003, p. 61), how long it takes for that correction to occur also dictates how long that window remains open for an investor to act on that information, which has yet to be processed by the market.

Another core tenet of the EMH that is often targeted is the notion of investors agreeing on the price implications of the information they have on hand. However, investors often behave irrationally and are seldom in agreement. If all investors in the stock market believe they are above average (Odean as cited in Kahneman, 2003b), it would be impossible for them to agree on the price of an asset, therefore making the market inefficient. In other words, Kahneman would argue that investors "are 'fully rational, except for…' some particular deviation that

explains a family of anomalies" (2003b, p. 163). As a result, the consideration of behavioural psychology sheds light on what influences investors, and by extension the stock market as well.

Specifically, behavioural psychologists, Kahneman and Tversky apply PT, which illustrates an investor's decision-making process in terms of framing and valuation processing (Tversky & Khaneman, 1992). This is to say that investors are swayed more by comparative measures, namely how much is gained or lost based on a reference point (Tversky & Khaneman, 1992). Ultimately, their findings show that there is a lot of emotional and subjective interpretation of information, down to simply "liking and disliking in factual predictions," which "indicate that traditional separation between belief and preference in analyses of decision making is psychologically unrealistic" (Kahneman, 2003a, p. 1470). Therefore, because an investor's decisions are subjective and inefficient in processing information, the market as a whole must be inefficient at times as well.

Sentiment Analysis

Now, bridging the gap between an efficient market and an inefficient market is the use of sentiment analysis. Regarding the similarities between Fama and Kahneman, both stress the importance information has on the stock market. In distilled form, Fama in particular stresses the accessibility of information and who has access to it, whereas Kahneman focuses more on how said information is interpreted and the psychological underpinnings of those decisions. Combining these theories, the abstract concepts of investor sentiment must be quantified so they may be objectively tested for their ability to predict market performance.

To provide a brief historical overview, the use of sentiment analysis in a financial context is a relatively recent development. Here, many of the sources consulted on sentiment analysis draw on research conducted by Baker and Wurgler who assert that fluctuations in sentiment are

predictive of fluctuations in the stock market (2006, 2007). However, as previously mentioned, the technology that allows for measuring sentiment through textual data is relatively new. As a result, the early work of Baker and Wurgler relied on financial "proxies" of sentiment, which include "the closed-end fund discount, NYSE share turnover, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium" (2006, p. 1655). Through these proxies Baker and Wurgler found the interpretation of sentiment to be a useful tool in predicting the stock prices of companies that were previously difficult to value through traditional methods, a conclusion that has since been further supported using non-proxy data as well (2007; Hribar & McInnis, 2012).

In relation to the EMH, sentiment analysis opens up a discussion on whether public sentiment is a "semi-strong form" or can be considered a "strong form." The initial thought would be to consider sentiment "semi-strong" because that information exists in the public domain. However, given the difficulty in acquiring the data and challenges in extracting meaningful information from the raw text an argument could be made for considering it a "strong form." This opens up the possibility for insight collected through sentiment analysis to not be incorporated by the market, hence hiding in plain sight, and thus presenting an informed investment opportunity.

Since the findings of Baker and Wurgler, textual data has been used to study the sentiment of both amateur and professional investors. Findings on amateur investors have found evidence that peer-based financial discussion on blogging sites and their comment sections can be used to predict stock market performance (Chen, De, Hu, & Hwang, 2014). Meanwhile, other studies have found that professional market analysts are affected by investor sentiment, which functions as an exogenous force that influences stock market fluctuations (Kaplanski & Levy,

2017). Research has also been critical on the actionability of these interpretations and found that by assigning a numerical value to text sources (Ranco et al., 2016), using Thermal Optimal Path calculations (Guo, Sun, & Qian, 2017), or layered attributes (Li, Chan, Ou, & Ruifeng, 2017), the actionability and accuracy of predictions increases.

Regarding some of the conclusions that have been made using sentiment analysis, a few studies have found that negative sentiment is a better predictor of financial loss than positive sentiment is of financial gain (Boudt & Petitjean, 2014; Tetlock, 2007). This finding also aligns with Kahneman's application of "loss-aversion" where "the response to losses is consistently much more intense than the response to corresponding gains, with a sharp kink in the value function at the reference point" (Kahneman, 2003b, p. 164). Combined, the findings of these studies support one another and point towards the possibility of using sentiment analysis in order to make predictions on future price fluctuations in the stock market.

Another finding worth mentioning is the notion that speculative stocks are more prone to being influenced by investor sentiment. As mentioned earlier, Timmermann and Granger found that in a financial context speculative assets are unforecastable. However, Baker and Wurgler found that opposite to be true as speculative assets, or "companies that are younger, smaller, more volatile, unprofitable, non-dividend paying, distressed, or with extreme growth potential" were more susceptible to shifts in investor sentiment (2007). These findings are also supported by Hribar and McInnis's study, which found when analysts forecast the future value of speculative assets they are more optimistic when prevailing sentiment is positive and less optimistic when prevailing sentiment is negative (2012).

The last point to mention on the subject of sentiment analysis involves discussing the literature of how sentiment is measured. Typically, the studies surveyed used a sentiment dictionary with words already coded to specific factors. Examples of such dictionaries include the Loughran–McDonald financial sentiment dictionary (Loughran & McDonald, 2011; Garcia, Chen, De, Hu, & Hwang, 2014; Li, Xie, Chen, Wang, & Deng, 2014; Loughran & McDonald, 2016), Baker and Wurgler's (2006) Sentiment Index (Hribar & McInnis, 2012; Kaplanski & Levy, 2017), or other independently designed corpuses (Guo, Sun, & Qian, 2017; Seng & Yang, 2017). Typically the implementation of these dictionaries involves tabulating the occurrences of coded words to provide a score for each sentiment category of the dictionary, which is then analyzed. For example, the RStudio version of the Loughran–McDonald financial sentiment dictionary codes words into six sentiments: Constraining, Litigious, Negative, Positive, Superfluous, and Uncertainty. However, most dictionaries are not all-encompassing, meaning not all words in the English language are coded. This means it is best to utilize a dictionary that has been made to analyze the sentiment of words typically found in the corpus being studied. Not doing so results in less fruitful data, and can lead to the misclassification of words, which harms the validity of any conclusions drawn from the analysis (Loughran & McDonald, 2011).

Machine Learning

Once the text has been coded for sentiment, it can then be used to make predictions. There are a variety of ways to do this but given that most of the studies surveyed in this literature review have large data sets it is prudent to utilize some form of algorithm to systematically make those predictions. One such algorithm is the Random Forest model developed by Leo Breiman in 2001, which is used for predicting categorical classifications. Currently, the Random Forest model has been applied in environmental studies to predict groundwater potential (Naghibi,

Pourghasemi, & Dixon, 2016), healthcare to improve identification of diseases (Mathotaarachchi et al., 2017), cybersecurity to recognize phishing emails (Akinyelu & Adewumi, 2014), and only recently in finance to predict stock movements (Weng, Lu, Wang, Megahed, & Martinez, 2018; Zhang, Cui, Xu, Li, & Li, 2018).

In terms of how the model works, the Random Forest generates a large number of independent decision trees to analyze the data and each tree makes a prediction as to which class it thinks the data corresponds to. The final prediction made by each tree is considered a vote and the class with the most votes is one the model ultimately predicts (Breiman, 2001). In a way, the Random Forest model's democratic process of determined predictions relates to one of the logical assumptions this study relies on, which is that on average the consensus of a group is the most accurate. In practice, this logical assumption asserts that the majority consensus among investor sentiment will also predict the direction of the stock market.

Another benefit of the Random Forest is that it prevents overfitting, which is detrimental to predictive accuracy. Essentially, overfitting results from a model that matches the training data too closely, and because of this, the model cannot accurately interpret new unseen data. In testing, this can be identified by a high training accuracy and a low testing accuracy. The Random Forest model through the Strong Law of Large Numbers shows that the predictions made by using large quantities of data and numerous decision trees "always converge so that overfitting is not a problem" (Breiman, 2001, p. 6).

Finally, a review of commonly used machine learning algorithms found the use of unigrams versus use of unigrams, bigrams, and trigrams combined, result in equal prediction accuracies (Pranckevičius, & Marcinkevičius, 2017). This is somewhat surprising, as one would assume that because bigrams and trigrams contain more information they would lead to more

accurate predictions. However, it is possible that the addition of more words leads to more noise, thus decreasing predictive accuracy. Moreover, very few sentences are composed of just two or three words, so bigrams and trigrams still cannot match the level of context contained in something like a sentence or paragraph, which are too large and specific to analyze. In addition, combinations of words also hinder the ability to use sentiment analysis, as combinations of words would result in combinations of sentiments. This dramatically increases the number of variables being interpreted and obfuscates the conclusions that can be drawn from them.

Research Questions

This study will analyze the corporate press releases of speculative and non-speculative stocks using the Loughran McDonald dictionary for financial sentiment. The purpose of this is to determine if the financial sentiment contained in press releases can predict daily fluctuations in stock prices. In addition, this study aims to uncover what sentiment or sentiments are most valuable in predicting stock price fluctuations, and whether or not those sentiments differ depending on the type of stock. Therefore, the research questions for this study are as follows:

RQ1. Can sentiment analysis of corporate press releases be used to predict negative and positive fluctuations in the stock market?

RQ2. Is there a difference in predictive accuracy using sentiment for speculative versus non-speculative stocks?

RQ3. Is there an optimal timeframe for predictive accuracy?

RQ4. What sentiments, contribute most to predictive accuracy?

RQ5. Do those sentiments differ based on a stock's level of speculation?

Methods

This section has been divided into seven phases with a reflection at the end. The phases have been organized in chronological order and detail the tasks and rationale behind each decision that was made. In sum, the phases involve discussing: the selection process for the stocks, the systematic approach for gathering textual data, processing the text, coding the text for sentiment, labeling each article for financial performance over time, building the machine learning model, and finally running statistical tests to determine which sentiments were significant. Given that this was a multi-phase endeavour, the reflection will discuss the elements that could be improved in future studies.

Phase One: Selecting the Stocks and Text Source

Selecting the Stocks

As touched on in the literature review, stocks can be divided into two categories, speculative and non-speculative. The notion of speculation, which Baker and Wurgler define as ease of valuation, can be judged based on "earnings history, tangible assets, and stable dividends," as well as volatility, profitability, and growth potential (2007, p. 132). These factors in mind, the speculative stocks selected were Aurora Cannabis (ACB), Canopy Growth Corporation (WEED), and Cronos Group (CRON). Meanwhile, the non-speculative stocks selected were Royal Bank of Canada (RY), Toronto–Dominion Bank (TD), and Scotiabank (BNS). The speculative stocks selected were all relatively new, highly volatile, non-dividend paying, currently unprofitable, and operate in a growth industry. Meanwhile, the non-speculative stocks were all long-established companies with an extensive earnings history, low volatility, dividend-paying, profitable and operate in a mature industry.

These stocks were selected as they had the highest market capitalization relative to their sectors. The rationale here was based on the assumption that the higher the market capitalization, the more coverage a stock would have, which is an important factor that will be covered again during the discussion section. These stocks were also selected to mitigate confounding variables. The speculative stocks not only all belong to the Life Sciences sector of the Toronto Stock Exchange (TSX), they also all belong to the same sub-sector, Cannabis. Likewise, the non-speculative stocks all belong to the Financial Services sector, and sub-sector, Banking.

Selecting the Text Source

Corporate press releases were chosen because they are the piece of communication that theoretically makes private information public information. As discussed in the literature review regarding the EMH, the moment when private information becomes public information is extremely important and the press release theoretically represents that moment. By analyzing the sentiment contained at the moment private information becomes public, this study may find a higher predictive accuracy than one would get studying a source that is based on information that has already been incorporated into stock's price. If the predictive accuracy is not higher in press releases, then this may suggest that sentiment contained in the commentary surrounding public information is more influential. In other words, information straight from the source may not be as influential of stock prices as the interpretation of said information by other influential figures.

Press releases were also chosen as they are linked to significant events, some reoccurring and others more novel. The reoccurring events reported in press releases were typically the announcement of quarterly results. Meanwhile, novel events reported in press releases could include information about class action lawsuits, new acquisitions, or other corporate initiatives that investors may find interesting. In order to capture a complete cycle of reoccurring

information as well as the news regarding more novel events, press releases were collected for the entire 2018-calendar year.

For the articles contained in the literature review, the actual source of textual data came from social media sites like Twitter (Li, Chan, Ou, & Ruifeng, 2017), and popular blogging sites such as Seeking Alpha (Chen, De, Hu, & Hwang, 2014), StockTwits (Li, Chan, Ou, & Ruifeng), and Yahoo! Finance (Ranco et al., 2016). However, this study's source for textual data was Marketwatch.com as they had a complete collection of archived press releases for each of the companies selected, and for the entire 2018-calendar year. There were other websites that also archived press releases, such as Yahoo! Finance and Reuters, but their collections were not complete and press releases more than six-months old were either deleted or slowly phased out. As a result, Marketwatch.com proved to be the best source for gathering all the textual information for this study.

Phase Two: Web Scraping

The process of web scraping can be defined as a systematically extracting and storing large quantities of data from websites. This was done using the RStudio packages rvest (Wickham, 2019a), stringr (Wickham, 2019b), and lubridate (Grolemund, & Wickham, 2011). The rvest package allowed text to be scraped by feeding in a search page's URL, identifying the desired information by plugging in its CSS selector, and then creating a "For Loop" to run through the search page and follow links to the articles so that the text could be scraped. Once scraped the lubridate package was used to ensure that all the articles had a uniform date code, which will become important later in the study. Finally, the stringr package was used to remove any unnecessary spaces before the scraped text was written to a CSV file and exported. This process was run for each stock until all the press releases for 2018 were collected.

The alternative to web scraping is manually selecting press releases to be studied. However, since this study aimed to utilize this information to make predictions, having as much information as possible was a priority. Furthermore, by collecting all the press releases for 2018, this study avoided the issue of accidentally selecting a sample that was biased or prone to some extraneous confounding variable.

Phase Three: Removing Unnecessary Text

Once the press releases were scraped and stored as a CSV, they needed to undergo a text cleaning process to remove words that may interfere with the coding for sentiment analysis. The press releases that were collected often contained biographical information about the company or press release agency, legal disclaimers, and contact information. For example, listed below are a few sample statements that would have been removed within the Scotiabank dataset:

Biographical Information About Scotiabank

> About Scotiabank Scotiabank is Canada's international bank and a leading financial services provider in the Americas. We are dedicated to helping our more than 25 million customers become better off through a broad range of advice, products and services, including personal and commercial banking, wealth management and private banking, corporate and investment banking, and capital markets. With a team of more than 97,000 employees and assets of $998 billion (as at October 31, 2018), Scotiabank trades on the Toronto Stock Exchange (BNS) and New York Stock Exchange BNS, -0.30% For more information, please visit www.scotiabank.com and follow us on Twitter @ScotiabankViews.

Information About the Press Release Agency

> About Investor Network Investor Network (IN) is a financial content community, serving millions of unique investors market information, earnings, commentary and news on the what's trending. Dedicated to both the professional and the average traders, IN offers timely, trusted and relevant financial information for virtually every investor. IN is an Issuer Direct brand, to learn more or for the latest financial news and market information, visit www.investornetwork.com. Follow us on Twitter @investornetwork. SOURCE: Investor Network https://www.accesswire.com/img.ashx?id=510766 Copyright 2018 ACCESSWIRE

Legal Disclaimer

> LEGAL NOTICES Information contained herein is not an offer or solicitation to buy, hold, or sell any security. Fundamental Markets, Fundamental Markets members, and/or Fundamental Markets affiliates are not responsible for any gains or losses that result from the opinions expressed. Fundamental Markets makes no representations as to the completeness, accuracy, or timeliness of the material provided and all materials are subject to change without notice. Fundamental Markets has not been compensated for the publication of this press release by any of the above mentioned companies. Fundamental Markets is not a financial advisory firm, investment adviser, or broker-dealer, and does not undertake any activities that would require such registration. For our full disclaimer, disclosure, and terms of service please visit our website. Media Contact: Andrew Duffie, Media Department Office: +1 667-401-0010 E-mail: media@Fundamental-Markets.com _—© 2018 Fundamental Markets. All Rights Reserved. For republishing permissions, please contact a partner network manager at partnership@Fundamental-Markets.com. CFA(R) and Chartered Financial Analyst(R) are registered trademarks owned by CFA Institute. FINRA(R), BrokerCheck(R), and CRD(R) are registered trademarks owned by Financial Industry Regulatory Authority, Inc. Copyright (C) 2018 GlobeNewswire, Inc. All rights reserved.

As illustrated in the example above, this text interferes with analysis by skewing the word count, which the sentiment analysis phase depends on. For example, the biographical information written by the company and for the company tends to include lots of positive words, which would positively skew the sentiment analysis results. Similarly, the legal disclaimers contain numerous litigious, constraining, and some negative words, which again would further skew results. Furthermore, this type of text appears in almost all the press releases across all the stocks that were selected. To remedy this, these pieces of text were manually removed. Unfortunately, this could not be automated at this point in the data collection process, which will be discussed in greater detail during the reflection on the methodology section.

Phase Four: Financial Coding of Articles

After collecting the text and removing biographical information and legal disclaimers, each press release was labelled based on the financial performance of the corresponding stock. In the literature review the most common sources for collecting financial data was the Institutional

Brokers' Estimate System (IBES) and Centre for Research in Securities Prices (CRSP) (Baker & Wurgler, 2006, 2007; Chen, De, Hu, & Hwang, 2014; Hribar & McInnis, 2012; Kaplanski & Levy, 2017). Fortunately, the financial information required for this research project was general enough to not require purchasing a licence or subscription to a database. Instead, the financial data for this project was collected by downloading Yahoo! Finance's historical data report for each stock.

For every trading day in 2018, the daily change in price was calculated by subtracting the opening price from the closing price that day. If the number were positive the trading day would be labelled "positive," and if the difference were negative the day would be labelled "negative." Then, to measure sentiment's ability to predict stock fluctuations over time, the closing day used to calculate the change in price would be pushed back 24-hours each time until a 96-hour delay was reached. The result was a table with every trading day in 2018 and a corresponding label (positive or negative) for each timeframe (same-day, 24-hours, 48-hours, 72-hours, and 96-hours). This also meant that in order to determine the label for dates at the end of 2018, stock data from the beginning of 2019 needed to be used.

Initially, a third "neutral" class was considered for labelling the textual content. In this scenario, the standard deviation of the stock movement would be calculated for each stock. If the movement on any day were greater than or equal to one standard deviation then it would be classified as "positive." Likewise, if the change were less than or equal to one standard deviation the stock's movement would be labelled as "negative." If the change in price did not exceed one standard deviation in either direction, the article would have been classified as "neutral." However, with machine learning, class imbalance dramatically skews results. For example, if the data set were 15% positive, 70% neutral, and 15% negative then the machine learning algorithm

could predict the neutral label for all the articles and have a 70% overall accuracy. At first glance, this seems like a good score. However, in reality this means the algorithm had no ability to make any meaningful predictions. Therefore, in order to avoid this issue a classification system with more symmetry was selected.

Phase Five: Sentiment Analysis

Tokenization and Stopword Removal

In order to conduct the sentiment analysis, the articles needed to be tokenized, which is the process of splitting the article into its constituent words. This phase primarily used the RStudio packages dplyr (Wickham, François, Henry, & Müller, 2019), and tidytext (Silge & Robinson, 2016). Once the text was tokenized, the stopwords were removed. This filters out words that often carry either very little or no sentiment value, for example: "a," "do," "for," "so," "the," and so on.

Sentiment Analysis

The Loughran–McDonald dictionary for financial sentiment was used to code the tokenized articles that were scraped. Conveniently, the dictionary was accessible through the tidytext package, and the articles were coded by date. The output of this was a table for every stock that contained a column for the timeframe and financial label, followed by columns containing the counts for each of the five sentiments (constraining, litigious, negative, positive, and uncertainty). It should be noted that the RStudio package also contained a sixth sentiment, "superfluous." While this sentiment was collected, it was ultimately ignored during the analysis as it contained words that were also considered stopwords. As a result, many of the superfluous

words were filtered out in the step prior. Furthermore, this also meant that some of the stocks had no count for this sentiment, making the comparison of this sentiment an unnecessary challenge.

Finally, the articles were coded for sentiment based on date. This created a representation of all the sentiments over time. Then, the table was exported as a CSV file so that the financial performance labels generated in phase four could be attached by date as well. If the articles were published on a non-trading day, the next trading day's labels were used.

Phase Six: Machine Learning Prediction Model

For this phase the RStudio packages primarily used were caret (Kuhn, 2019) and randomForest (Liaw & Wiener 2002). First, the financially labelled and sentiment coded stock information was imported into RStudio. Then, the financial label being tested (either same-day, 24-hour delay, 48-hour delay, 72-hour delay, or 96-hour delay) was converted into a factor variable. Following that step, the caret package was used to create a random stratified 70/30 split to maintain the proportion of positive and negative financial labels between the test (30% portion) and training (70% portion) sets. The unused financial label columns were then nulled to maintain some clarity in the data.

After preparing the text the randomForest package was implemented to execute the Random Forest algorithm for machine learning on the training data. Once completed, the test data was inputted for the algorithm to make its predictions. To interpret these predictions a confusion matrix was used to determine the balanced accuracy, sensitivity (accuracy at predicting the negative label) and specificity (accuracy at predicting the positive label). Last, the randomForest package also had a function that could be used to determine how many times each sentiment was used by the Random Forest model when making its prediction.

For recording the results, the machine learning algorithm was run five times for each time frame as each run yielded different results. This randomness occurs in part due to having different articles contained the training and test data sets each time the random stratified split is done. In turn, the Random Forest model yields different prediction accuracies as it is trained using slightly different data each time. Furthermore, the Random Forest model is a stochastic system, which depends on randomly generated decision trees. Combined, these two factors are what lead to variation in the test accuracy results.

The variation of results can be avoided by setting a random seed to generate the same random data split each time. However, in a production setting, this is not advised as it would mean hand-picking a seed that produces the highest test accuracy, which may not necessarily be the highest accuracy when new unseen data is introduced. As a result, the balanced accuracy, sensitivity, specificity, and sentiment use scores were recorded across five trials at each timeframe in order to capture the variation that results from having random stratified splits of data.

Phase Seven: Regression and Pearson Correlation Analysis

Finally, after gathering the results from the machine learning phase a regression and Pearson's correlation analysis were conducted. This was done to determine which sentiment or sentiments were statistically significant in the prediction accuracy of speculative versus non-speculative stocks, and in prediction accuracy over time. To answer the speculative versus non-speculative question, the results of the machine learning phase were sorted by type of stock. Last, to answer the predictive accuracy over time all the trial scores for each timeframe were combined and analyzed.

Reflection on Methodology

In a perfect world, every step in this methods section would have been completed without intervention, and while the majority of tasks were completed this way, some phases did require manual input. In phase two, the web scraping required each page to be manually setup for scraping. Then, to move through pages the customized URL had to be entered into the code for every page of search results in order to scrape all the press releases. In a future study, it might be wise to create a function that loops through the URL by their pattern so that one could move through all the desired pages more efficiently once they have been setup.

Moreover, phase three could have been amalgamated into phase two had a slightly different scraping method been chosen. Instead of scraping all the paragraph tags and then collapsing them into one cell, each paragraph tag could have been scraped into its own cell and stored that way. If the scraping were done like this, then a library of biographical information and legal disclaimers could have been made and used to systematically remove all extraneous text.

Apart from improving the automation of the web scraping, the sentiment analysis phase also presents some opportunity for improvement. While the Loughran–McDonald dictionary for financial sentiment is purpose-built for interpreting financial texts, it does lose out on some of the more novel words that may be included in a press release. For example, if a company acquires another company, chances are the acquired company's name will be excluded from analysis even though it could potentially add a lot of predictive value. In response to this, a separate dictionary could be constructed and added as an additional sentiment to the ones already contained in the Loughran–McDonald dictionary.

Finally, regarding the machine learning phase, the randomForest package allows the algorithm to be tuned and customized based on test results. To keep things simple when comparing results, no tuning measures were taken. As a result, the accuracy scores are not as high as they could be. If this algorithm were to be used in a production setting then the algorithm would be tuned for each stock to try and maximize for predictive accuracy. However, this level of intervention would have gone beyond the scope of this study, as it would have introduced factors that would have made the comparison of results more difficult.

In sum, this current structure for the methods section was successful in systematically acquiring, processing, labelling, coding, testing, and analyzing the textual information. The only elements that one might want to improve upon for future studies would be increasing the efficiency in executing some of the processing tasks, improving the granularity of the sentiment analysis, and perhaps refining the machine learning parameters.

Results

The results of each phase are listed in the order that they were collected. This section will outline the basic details of the data collected, the results of the sentiment analysis, the accuracy scores from running the machine learning algorithm, and the results of the regression and Pearson analysis.

Summary of Data Collected

Table 1 provides a summary of the press releases that were collected through the web scraping process. The information has been organized by type of stock (non-speculative and speculative) and is further divided into each specific stock. Also included in this table is a break down of how the articles were labelled across the various timeframes (Same Day, 24-Hours, 48-Hours, 72-Hours, and 96-Hours). The purpose of this table is to help provide context for the discussion to follow.

Table 1: Summary of Scraped Press Releases

Stock Type	Stock	Number of Articles	Same Day	24-Hours	48-Hours	72-Hours	96-Hours
			Number of Positive Labels / Number of Negative Labels				
Non-Speculative	BNS	29	9/20	7/22	8/21	8/21	6/23
	RY	30	16/14	17/13	17/13	17/13	18/12
	TD	54	26/28	27/27	26/28	31/23	27/27
Subtotal of Non-Speculative		**113**	**51/62**	**51/62**	**51/62**	**56/57**	**51/62**
Speculative	ACB	483	193/290	211/272	201/282	197/286	198/285
	CRON	261	110/151	106/155	108/153	113/148	108/153
	WEED	334	149/185	139/195	147/187	148/186	151/183
Subtotal of Speculative		**1078**	**452/626**	**456/622**	**456/622**	**458/620**	**457/621**
Total Press Releases		**1191**	**503/688**	**507/684**	**507/684**	**514/677**	**508/683**

Sentiment Analysis Scores

Tables 2.1 to 2.6 and figures 1.1 to 1.6 provide an in-depth view of the sentiment scores for each stock using the Loughran–McDonald dictionary for financial sentiment. Specifically, the tables were generated to summarize the sentiment scores of each stock across the five time frames. This allows for interpretation of how sentiment changes when considering price fluctuations over various lengths of time.

Meanwhile, the figures provide a visualization of what words were used most often in the stock's press releases for each sentiment. This visualization also allows patterns in the words used to be identified. For example, words which appear most often among the same type of stocks, and/or across the types of stocks.

From the table, it becomes apparent that the two most common sentiments by count are Negative and Positive. However, this likely has more to do with there being more words coded under these sentiments than any other.

Table 2.1: Scotiabank Sentiment Counts by Stock Performance

Timeframe	Performance	Sentiment Counts by Stock Performance					
		Constraining	Litigious	Negative	Positive	Superfluous	Uncertainty
Same Day	Negative	2	1	12	15	0	3
	Positive	3	1	23	25	0	3
24-Hours	Negative	2	0	16	26	0	4
	Positive	3	2	17	13	0	3
48-Hours	Negative	2	1	16	27	0	2
	Positive	3	1	17	14	0	3
72-Hours	Negative	2	1	16	27	0	4
	Positive	3	1	17	14	0	3
96-Hours	Negative	2	1	25	27	0	5
	Positive	3	1	6	14	0	1

Figure 1.1: Scotiabank Most Used Words by Sentiment

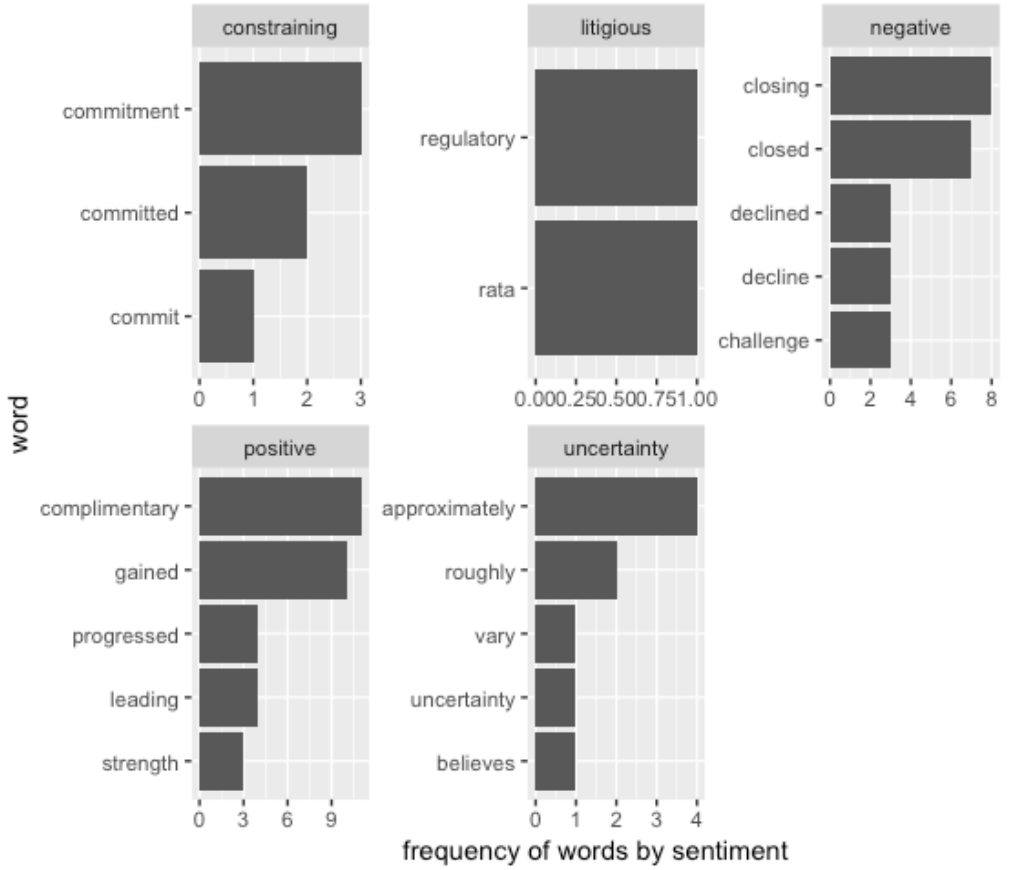


Table 2.2: Royal Bank Sentiment Counts by Stock Performance

Timeframe	Performance	Sentiment Counts by Stock Performance					
		Constraining	Litigious	Negative	Positive	Superfluous	Uncertainty
Same Day	Negative	5	1	15	24	0	3
	Positive	2	1	6	19	0	3
24-Hours	Negative	4	1	14	23	0	3
	Positive	3	1	8	22	0	3
48-Hours	Negative	4	1	10	22	0	3
	Positive	3	1	12	23	0	3
72-Hours	Negative	4	0	14	21	0	3
	Positive	3	2	8	23	0	3
96-Hours	Negative	4	1	14	23	0	3
	Positive	3	1	8	22	0	3

Figure 1.2: Royal Bank Most Used Words by Sentiment

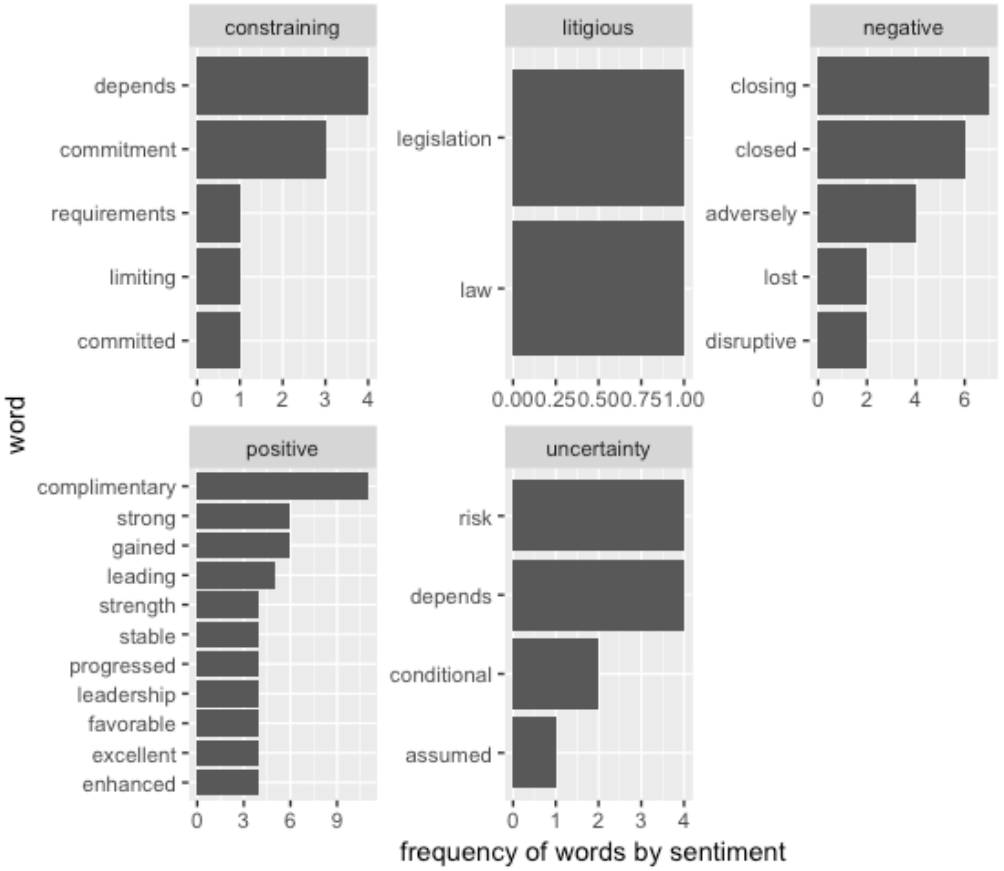


Table 2.3: Toronto Dominion Bank Sentiment Counts by Stock Performance

Timeframe	Performance	Sentiment Counts by Stock Performance					
		Constraining	Litigious	Negative	Positive	Superfluous	Uncertainty
Same Day	Negative	6	15	57	81	1	16
	Positive	5	8	42	89	0	9
24-Hours	Negative	7	9	53	90	1	13
	Positive	5	10	44	84	0	12
48-Hours	Negative	7	9	58	97	1	13
	Positive	5	10	38	72	0	10
72-Hours	Negative	7	16	50	87	0	16
	Positive	5	5	48	83	1	8
96-Hours	Negative	7	16	51	87	0	16
	Positive	5	5	47	83	1	8

Figure 1.3: Toronto Dominion Bank Most Used Words by Sentiment

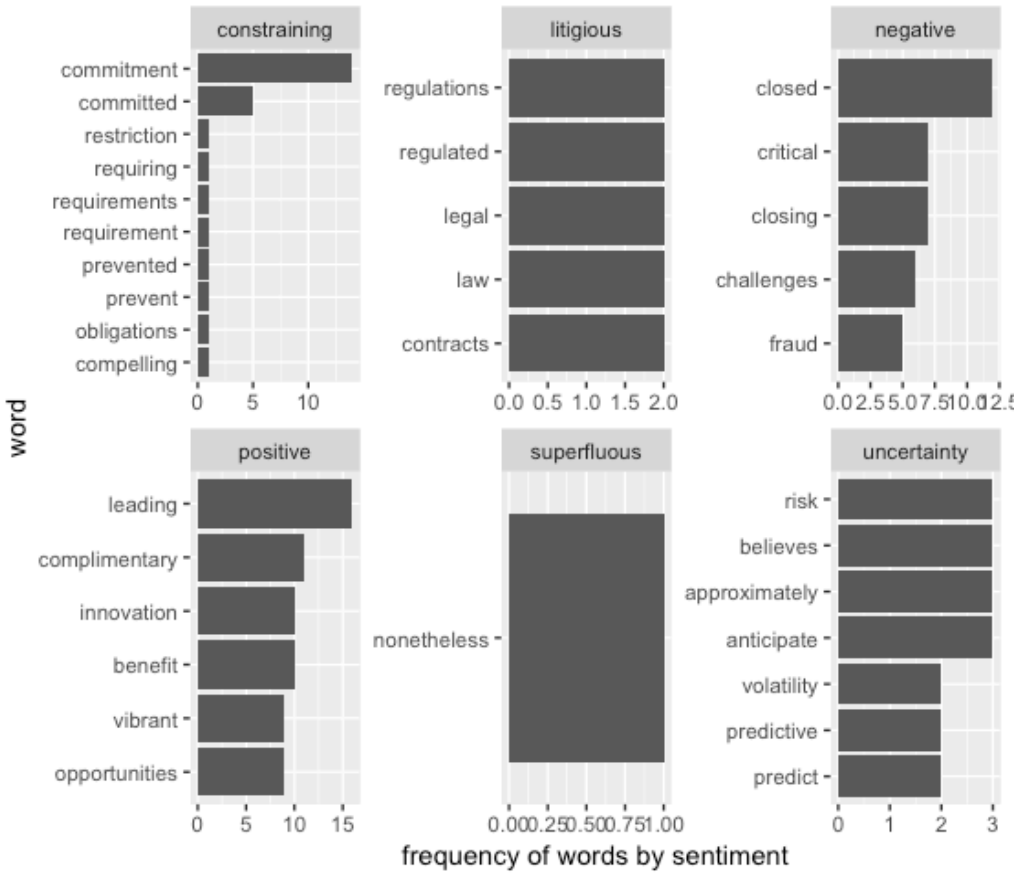


Table 2.4: Aurora Cannabis Sentiment Counts by Stock Performance

Timeframe	Performance	Sentiment Counts by Stock Performance					
		Constraining	Litigious	Negative	Positive	Superfluous	Uncertainty
Same Day	Negative	56	73	286	205	2	85
	Positive	40	60	179	180	5	65
24-Hours	Negative	54	71	289	210	3	85
	Positive	43	58	181	177	4	66
48-Hours	Negative	55	63	283	212	4	84
	Positive	42	64	190	166	4	63
72-Hours	Negative	52	66	279	206	4	76
	Positive	48	62	189	180	4	69
96-Hours	Negative	51	59	266	207	4	75
	Positive	48	67	208	179	4	76

Figure 1.4: Aurora Cannabis Most Used Words by Sentiment

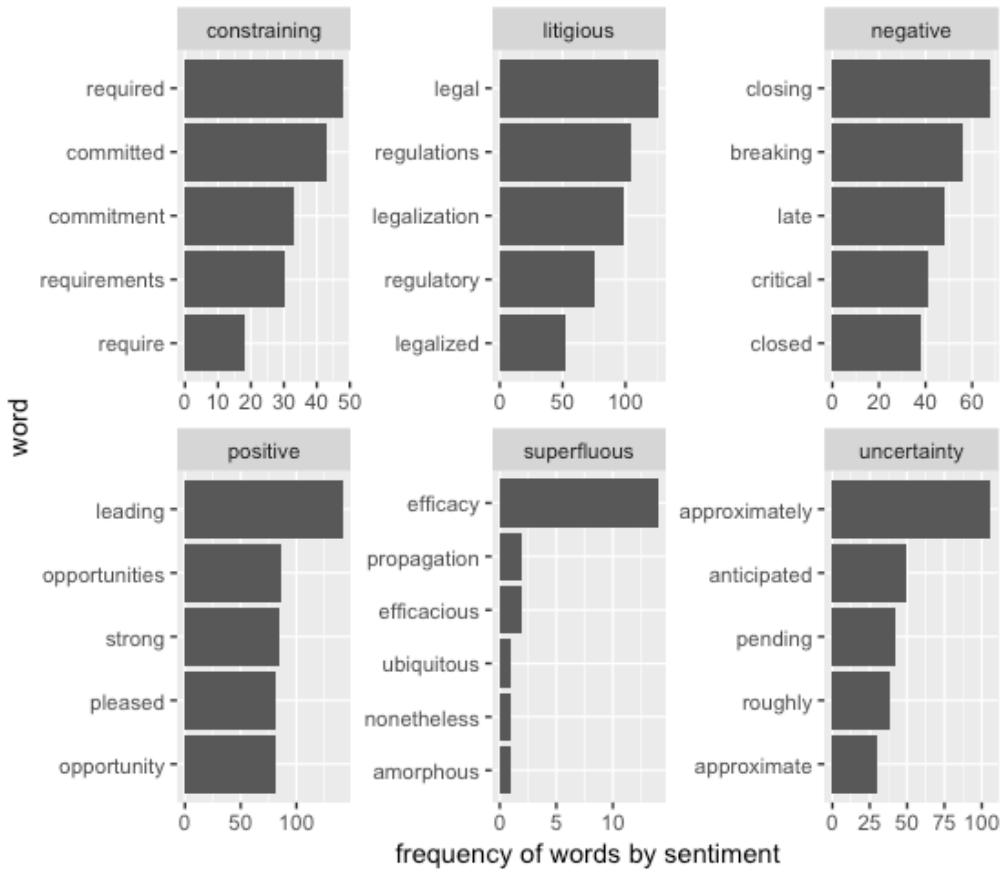


Table 2.5: Cronos Group Sentiment Counts by Stock Performance

Timeframe	Performance	Sentiment Counts by Stock Performance					
		Constraining	Litigious	Negative	Positive	Superfluous	Uncertainty
Same Day	Negative	40	69	196	172	5	61
	Positive	30	61	152	146	3	40
24-Hours	Negative	44	67	211	176	5	63
	Positive	24	62	119	141	4	36
48-Hours	Negative	43	67	209	174	5	61
	Positive	26	65	132	151	4	40
72-Hours	Negative	43	64	197	174	5	60
	Positive	27	63	141	152	4	40
96-Hours	Negative	44	65	200	179	5	61
	Positive	24	64	138	147	3	40

Figure 1.5: Cronos Group Most Used Words by Sentiment

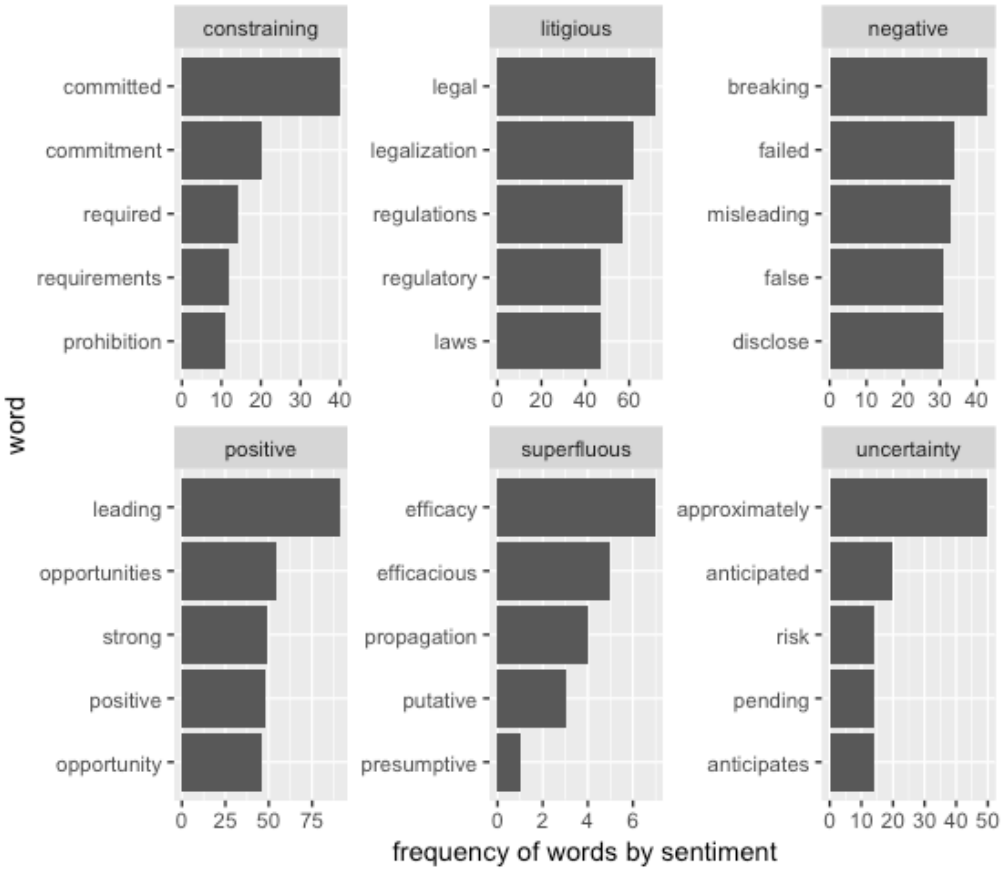
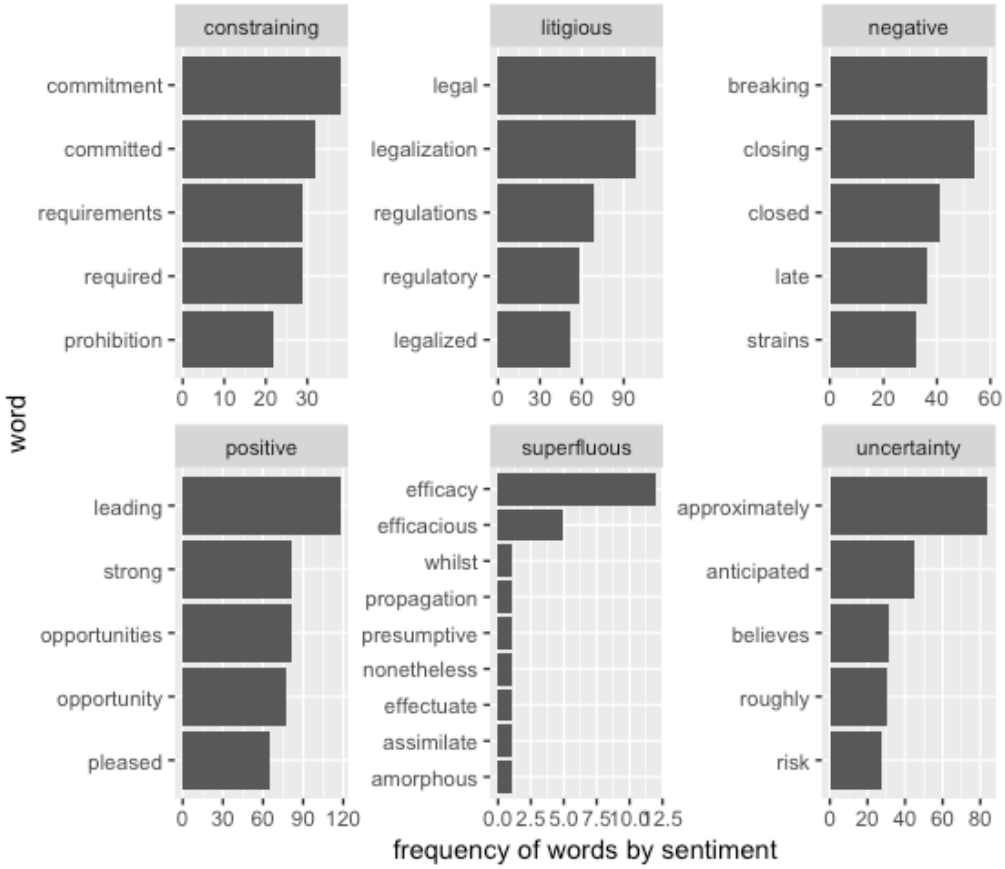


Table 2.6: Canopy Growth Corporation Sentiment Counts by Stock Performance

Timeframe	Performance	Sentiment Counts by Stock Performance					
		Constraining	Litigious	Negative	Positive	Superfluous	Uncertainty
Same Day	Negative	49	70	251	186	4	82
	Positive	40	61	193	188	7	69
24-Hours	Negative	49	69	253	193	8	85
	Positive	41	60	194	186	3	66
48-Hours	Negative	48	67	242	189	6	85
	Positive	41	60	194	185	5	61
72-Hours	Negative	48	68	245	190	6	88
	Positive	40	60	190	189	5	59
96-Hours	Negative	47	71	238	186	5	78
	Positive	44	63	208	195	6	74

Figure 1.6: Canopy Growth Corporation Most Used Words by Sentiment



Machine Learning Prediction Accuracy

After completing the machine learning trials for each stock and timeframe, the results were averaged and recorded in tables 3.1 and 3.2, non-speculative and speculative respectively. The purpose of these tables is to help answer research questions one to three. These questions pertain to: the possibility of using sentiment analysis of press releases to predict stock price fluctuations, whether or not the prediction accuracy varies depending on the amount of speculation, and whether or not there is an optimal time frame for predictive accuracy.

Meanwhile, tables 4.1 and 4.2 contain the average number of times each sentiment was used per timeframe during the machine learning process to make a prediction. Therefore, the purpose of these tables is to provide context for addressing research questions four and five. These questions seek to find out what sentiments contribute most to predictive accuracy, and whether or not predictive sentiments differ between based on the level of speculation.

The results of table 3.1 show that for non-speculative stocks, the highest average negative test accuracy is 62.41%, which appears in the 96-hour timeframe. Conversely, the highest average positive test accuracy recorded at 51.11% appears in the same-day timeframe. However, the table also shows that this variation in optimal timeframe is not as dramatic as it seems since the second highest predictive accuracy for negative movements, which is 60.79%, appears in the same-day timeframe. As a result, the highest balanced accuracy (55.95%) appears in the same-day timeframe as well.

Meanwhile, for speculative stocks, the results of table 3.2 show that the highest average negative test accuracy of 66.43% also appears in the 96-hour timeframe. Unlike the non-speculative stocks however, the highest average positive test accuracy of 44.73% appears in the 48-hour timeframe. The highest balanced accuracy of 55.38% also appears in the 48-hour

timeframe, as the second highest average negative test accuracy of 66.04% also appears in the 48-hour timeframe.

Regarding tables 4.1 and 4.2, any findings are difficult to ascertain simply by looking at the scores. For many of the sentiments, the differences in scores vary only marginally across timeframes. Furthermore, there is a disparity between the raw count of sentiments as the Loughran-McDonald dictionary for financial sentiment simply has more words coded for positive and negative sentiment. To better tease out results, the next phase employs regression analysis and a Pearson's correlation. For now however, tables 4.1 and 4.2 simply serve as a summary of the coded scores.

Table 3.1: Non-Speculative Stock Prediction Accuracy

Timeframe	Stocks	Test Accuracy		
		Negative	Positive	Balanced
Same Day	BNS	90.00%	20.00%	55.00%
	RY	46.67%	80.00%	63.34%
	TD	45.71%	53.33%	49.52%
	Average	**60.79%**	**51.11%**	**55.95%**
24-Hours	BNS	85.00%	10.00%	47.50%
	RY	20.00%	75.00%	47.50%
	TD	40.00%	53.33%	46.67%
	Average	**48.33%**	**46.11%**	**47.22%**
48-Hours	BNS	80.00%	0.00%	40.00%
	RY	26.67%	46.67%	36.67%
	TD	42.86%	26.67%	34.76%
	Average	**49.84%**	**24.45%**	**37.14%**
72-Hours	BNS	85.00%	10.00%	47.50%
	RY	20.00%	66.67%	43.33%
	TD	33.33%	43.33%	38.33%
	Average	**46.11%**	**40.00%**	**43.06%**
96-Hours	BNS	92.00%	0.00%	46.00%
	RY	26.67%	66.67%	46.67%
	TD	68.57%	36.00%	52.29%
	Average	**62.41%**	**34.22%**	**48.32%**

Table 3.2: Speculative Stock Prediction Accuracy

Timeframe	Stocks	Test Accuracy		
		Negative	Positive	Balanced
Same Day	ACB	75.56%	22.73%	49.14%
	CRON	54.55%	58.95%	56.75%
	WEED	61.48%	43.81%	52.65%
	Average	**63.86%**	**41.83%**	**52.84%**
24-Hours	ACB	60.57%	36.52%	48.55%
	CRON	66.09%	52.22%	59.16%
	WEED	59.23%	41.82%	50.52%
	Average	**61.96%**	**43.52%**	**52.74%**
48-Hours	ACB	73.89%	34.55%	54.22%
	CRON	59.05%	66.00%	62.52%
	WEED	65.19%	33.63%	49.41%
	Average	**66.04%**	**44.73%**	**55.38%**
72-Hours	ACB	73.51%	28.57%	51.04%
	CRON	58.09%	50.00%	54.05%
	WEED	62.22%	27.62%	44.92%
	Average	**64.61%**	**35.40%**	**50.00%**
96-Hours	ACB	74.45%	25.45%	49.95%
	CRON	58.18%	53.68%	55.93%
	WEED	66.67%	31.82%	49.24%
	Average	**66.43%**	**36.99%**	**51.71%**

Table 4.1: Sentiments Used to Make Predictions for Non-Speculative Stocks

Timeframe	Stocks	Sentiments Used to Make Predictions				
		Constraining	Litigious	Negative	Positive	Uncertainty
Same Day	BNS	282.20	118.40	711.40	678.80	330.00
	RY	319.60	117.60	596.60	544.20	299.80
	TD	586.80	349.40	1028.60	1213.00	600.60
	Average	**396.20**	**195.13**	**778.87**	**812.00**	**410.13**
24-Hours	BNS	241.40	235.40	550.40	617.80	322.80
	RY	374.00	153.60	662.40	631.40	340.00
	TD	630.00	430.40	1170.60	1290.00	699.60
	Average	**415.13**	**273.13**	**794.47**	**846.40**	**454.13**
48-Hours	BNS	258.00	119.20	772.80	694.80	349.20
	RY	445.40	180.80	630.80	732.60	386.40
	TD	593.20	554.00	1063.40	1179.80	623.80
	Average	**432.20**	**284.67**	**822.33**	**869.07**	**453.13**
72-Hours	BNS	220.00	159.20	755.80	742.20	405.60
	RY	318.20	130.40	630.80	605.20	378.60
	TD	699.40	442.20	1146.20	1277.00	726.60
	Average	**412.53**	**243.93**	**844.27**	**874.80**	**503.60**
96-Hours	BNS	277.20	135.60	648.60	676.20	285.60
	RY	443.00	171.80	680.60	704.60	346.80
	TD	637.20	420.60	1157.40	1248.20	690.40
	Average	**452.47**	**242.67**	**828.87**	**876.33**	**440.93**

Table 4.2: Sentiments Used to Make Predictions for Speculative Stocks

Timeframe	Stocks	Sentiments Used to Make Predictions				
		Constraining	Litigious	Negative	Positive	Uncertainty
Same Day	ACB	2897.80	3793.00	3996.20	4502.20	3435.20
	CRON	1976.00	3074.40	3087.20	3160.40	2169.20
	WEED	2561.80	3144.80	3421.00	3867.80	2821.40
	Average	**2478.53**	**3337.40**	**3501.47**	**3843.47**	**2808.60**
24-Hours	ACB	2815.80	3798.20	4032.40	4477.20	3384.00
	CRON	1797.80	2884.40	2801.20	2893.60	2026.00
	WEED	2660.20	3061.80	3448.80	3988.40	3002.00
	Average	**2424.60**	**3248.13**	**3427.47**	**3786.40**	**2804.00**
48-Hours	ACB	2746.80	3629.60	4000.40	4440.00	3362.40
	CRON	1868.80	2762.60	2775.00	2786.20	2039.80
	WEED	2607.60	3104.60	3458.20	3957.00	2877.20
	Average	**2407.73**	**3165.60**	**3411.20**	**3727.73**	**2759.80**
72-Hours	ACB	2848.60	3530.60	3967.20	4407.20	3394.40
	CRON	1910.40	2821.60	2735.20	2932.00	2021.60
	WEED	2689.60	3276.20	3624.60	4108.80	3085.60
	Average	**2482.87**	**3209.47**	**3442.33**	**3816.00**	**2833.87**
96-Hours	ACB	2713.80	3484.60	3759.20	4390.00	3509.00
	CRON	1892.20	2950.40	2804.40	2992.00	2029.80
	WEED	2573.20	3098.60	3387.80	3972.40	3053.60
	Average	**2393.07**	**3177.87**	**3317.13**	**3784.80**	**2864.13**

Statistical Tests: Regression Analysis and Pearson Correlation

Tables 5.1 and 5.2 display the results of the regression analysis, which are grouped by level of speculation, non-speculative and speculative respectively. In each table, the results highlight the sentiments that were statistically significant to the predictive accuracy of positive and negative stock price fluctuations, as well as what timeframe they were statistically significant in. Below each table are the significance codes, which provide insight into the sentiment's level of significance.

At a glance, tables 5.1 and 5.2 show that there are only a few statistically significant sentiments and potential timeframes. In table 5.1 (non-speculative), we see a mix of the constraining, litigious, and positive sentiments as statistically significant. The constraining sentiment appears to have been statistically significant most often, and significant at the highest level. Table 5.2 (speculative) on the other hand, shows no statistical significance for the constraining sentiment. Instead, there is a mix of the negative and uncertainty sentiments as statistically significant.

Tables 5.3 and 5.4 reveal the Pearson correlations for the non-speculative and speculative stocks respectively. This test was conducted to measure the relationship between the number of times a sentiment was used during the Random Forest model and the resulting predictive accuracy. The asterisk in these tables denotes the correlation coefficient with the most significant effect size by row, which relates to the timeframe and financial label.

In table 5.3 (non-speculative) the constraining sentiment almost always had the most significant effect size, with the exception of the one litigious sentiment. However, the effect sizes recorded here are only within the moderate range to minor range. Moreover, some of the effect sizes are negligible, indicating no relationship to predictive accuracy. The opposite however can

be seen in table 5.4 (speculative) as most of the significant effect sizes fall within the highest range. Furthermore, the strongest correlations by row were far more spread out as all five sentiments had the highest correlation coefficient by row at least once. There are however noticeable clusters that appear around the uncertainty, negative, and positive sentiments.

Table 5.1: Regression Analysis Results for Non-Speculative Stocks

Timeframe	Financial Label	P-Value				
		Constraining	Litigious	Negative	Positive	Uncertainty
Same Day	Negative	0.318	0.913	0.443	0.487	0.916
	Positive	0.865	0.671	0.581	0.507	0.586
24-Hours	Negative	0.166	0.601	0.534	0.960	0.986
	Positive	0.010 ***	0.903	0.569	0.933	0.420
48-Hours	Negative	0.736	0.385	0.978	0.572	0.295
	Positive	0.774	0.685	0.112	0.326	0.512
72-Hours	Negative	0.262	0.127	0.421	0.359	0.204
	Positive	0.003 ***	0.107	0.152	0.086 *	0.187
96-Hours	Negative	0.551	0.747	0.394	0.744	0.582
	Positive	0.103	0.013 **	0.760	0.067 *	0.699

Significance Codes:
*** 0.01 | ** 0.05 | * 0.1

Table 5.2: Regression Analysis Results for Speculative Stocks

Timeframe	Financial Label	P-Value				
		Constraining	Litigious	Negative	Positive	Uncertainty
Same Day	Negative	0.656	0.714	0.749	0.913	0.666
	Positive	0.802	0.263	0.659	0.211	0.047 **
24-Hours	Negative	0.113	0.571	0.695	0.836	0.526
	Positive	0.587	0.150	0.094 *	0.314	0.572
48-Hours	Negative	0.849	0.171	0.287	0.790	0.573
	Positive	0.974	0.269	0.481	0.166	0.317
72-Hours	Negative	0.910	0.329	0.016 **	0.709	0.051 *
	Positive	0.472	0.170	0.099 *	0.941	0.308
96-Hours	Negative	0.367	0.963	0.331	0.929	0.915
	Positive	0.552	0.273	0.554	0.980	0.394

Significance Codes:
*** 0.01 | ** 0.05 | * 0.1

Table 5.3: Pearson Correlation Results for Non-Speculative Stocks

Timeframe	Financial Label	Pearson Correlation Coefficient				
		Constraining	Litigious	Negative	Positive	Uncertainty
Same Day	Negative	-0.17	-0.34 *	-0.21	-0.19	-0.15
	Positive	-0.07	-0.01	-0.09*	-0.09*	-0.07
24-Hours	Negative	-0.39 *	0.06	-0.29	-0.27	-0.12
	Positive	0.36 *	-0.01	0.27	0.23	0.10
48-Hours	Negative	-0.40 *	-0.22	0.18	-0.10	-0.18
	Positive	0.34 *	0.14	-0.14	0.14	0.17
72-Hours	Negative	-0.46 *	-0.12	-0.01	-0.10	-0.17
	Positive	0.27 *	-0.01	-0.19	-0.06	-0.04
96-Hours	Negative	-0.24 *	0.12	0.08	0.12	0.05
	Positive	0.44 *	0.07	0.06	0.07	0.18

* Signals the sentiments with the greatest magnitude of correlation by financial label.

Table 5.4: Pearson Correlation Results for Speculative Stocks

Timeframe	Financial Label	Pearson Correlation Coefficient				
		Constraining	Litigious	Negative	Positive	Uncertainty
Same Day	Negative	0.66	0.72	0.73 *	0.70	0.71
	Positive	-0.80	-0.78	-0.76	-0.84 *	-0.79
24-Hours	Negative	-0.48 *	-0.09	-0.28	-0.34	-0.30
	Positive	-0.58	-0.56	-0.67 *	-0.62	-0.65
48-Hours	Negative	0.56	0.74 *	0.63	0.66	0.67
	Positive	-0.83	-0.28	0.93	0.92	0.97 *
72-Hours	Negative	0.58	0.71 *	0.66	0.60	0.54
	Positive	-0.75	-0.67	-0.79 *	-0.75	-0.73
96-Hours	Negative	0.50	0.53	0.65 *	0.60	0.56
	Positive	-0.69	-0.55	-0.77	-0.78 *	-0.78 *

* Signals the sentiments with the greatest magnitude of correlation by financial label.

Discussion

This section will unpack the results section, and draw some conclusions based on the findings. Beginning with the results gathered from the web scraping, this discussion will address some of the larger market forces at play, and how they may influence the results. Following this, the discussion section will work through answering the research questions by interpreting the findings from the results section and linking them to the theories mentioned in the literature review. Research questions one to three will be addressed through an examination of the machine learning while research questions four and five results will be addressed by interpreting the results of the regression analysis and Pearson correlation tests. Once the research questions have been addressed, this discussion section will use those findings in an attempt to isolate specific words and phrases to provide more insight. This will involve using the figures displaying the most used words by sentiment for each stock, and tracing them back to where they originated in the press releases. This will allow the context of the words to be considered as well. Finally, this section will conclude by discussing the limitations as well as areas of future study.

Properties of the Data Set

In total 1191 press releases were scraped for use in this study. Of those 1191, only 113 of them belonged to non-speculative stocks, with the remaining 1078 coming from the speculative stocks. This roughly ten to one imbalance in press release coverage might represent something that can be used to characterize the difference between speculative and non-speculative stocks. For example, in addition to Baker and Wurgler's use of company size, earnings history, volatility and so on to define what is or is not a speculative stock, the volume of press coverage can also indicate a stock's level of speculation. Logically, this imbalance in coverage makes sense as a general characteristic because new information pertaining to new stocks has the ability to

materially influence its value. This is because the EMH is constantly at play, correcting prices to reflect all public knowledge of that company. Part of that is considering long-term trends of the company, but if the company is new, it cannot have long-term trends, meaning presently available information is even more important and influential. This lack of history is what forces investors to speculate on new stocks, therefore the more information the better.

The other property worth discussing is the consistent class imbalance between positive and negative financial performance labels. The reason for this slant towards the negative is because the year 2018 as a whole was down. The S&P/TSX Composite Index fell roughly 2,300 points or a drop of approximately 15%. Coincidentally, this is also roughly the difference between the negative and positive labels generated for the financial performance of the press releases. This bias may reduce the accuracy in predicting positive fluctuations simply because there will be less training data, less test data, and possibly less influential positive sentiment if the prevailing market sentiment is broadly negative. That being said, class imbalances are inevitable especially given the unpredictable nature of the stock market. The remedy here would simply be to sample more stocks in order to increase the data set as a whole so that the number of positive labels increases objectively.

Examination of Machine Learning

Regarding research question one: can sentiment analysis of corporate press releases be used to predict negative and positive fluctuations in the stock market? The answer in short, would be yes, but it depends on the type of fluctuation and timeframe. The explanation here also ties in with research question two: is there a difference in predictive accuracy using sentiment for speculative versus non-speculative stocks? And the answer to this in short would also be yes, but it too depends on the timeframe and type of fluctuation.

To provide some context, this study considers an accuracy of less than 50% to be a failure, as the prediction would be statistically worse than a simple coin toss or random chance. As touched on during the results section, table 3.1 shows that for non-speculative stocks, the highest average balanced test accuracy appears in the same-day time frame with an accuracy of 55.95%. As for predictive accuracy of non-speculative stocks, table 3.2 shows that the highest average balanced accuracy appears in the 48-hour time frame with a value of 55.38%.

Since 55.95% and 55.38% are higher than 50%, this would be considered a success as the machine learning algorithm used sentiment analysis to provide an edge of 5.95% and 5.38% edge respectively. However, table 3.1 and 3.2 also illustrates that the overall accuracy is weighed down by consistently subpar accuracies in predicting positive movements. As discussed earlier this may stem from the fact that 2018 overall was not a positive year, which suggests that a down market hinders prediction accuracy of positive fluctuations. This might be because the setup of this study only looks at stock and sector-specific sentiment, while the broader market sentiment is not accounted for.

As for research question three: is there an optimal time frame for predictive accuracy? The answer would be yes again, however it depends on whether or not the sector is speculative or non-speculative. For non-speculative stocks the same-day time frame provides the highest level of accuracy, meanwhile for speculative stocks, it is the 48-hour time frame. The reason for this difference may come down to factors beyond sentiment. Established non-speculative stocks likely have established factors that have a bigger influence on price fluctuations, and given that those factors are well established they can be incorporated into the stocks price more efficiently. Less established and more speculative stocks however likely have less established factors that

influence the price. As a result of this inefficiency, these stocks take longer to incorporate these factors into their price.

Interpretation of Regression Analysis and Pearson Correlations

As for research questions four and five: what sentiments, contribute most to predictive accuracy? And, do those sentiments differ based on a stock's level of speculation? As shown in table 5.1, the regression analysis identified the constraining, litigious, and positive sentiments as statistically significant in the predictive accuracy of non-speculative stocks. Then, in table 5.3, the Pearson correlation further supported the results of the regression analysis as it identified the constraining sentiment as having the largest effect size of all the sentiments by row, with the exception of the one instance where it was the litigious sentiment, which still is in line with the regression analysis.

An interesting observation here is that the statistically significant sentiments all appear after the same-day time frame, despite the same-day timeframe having the highest average balanced accuracy score. The significant sentiments also only apply to predictions of positive fluctuations, which the machine learning model consistently had trouble predicting. As for the Pearson correlations, table 5.3 asserts that there is almost always a positive correlation between the constraining sentiment and positive prediction accuracy, and a negative correlation between the constraining sentiment and the negative prediction accuracy. When the results of these two tests are combined it would seem that the appearance of the constraining sentiment is influential across all time frames. However, on its own it is not influential enough to improve accuracy. To remedy this, the constraining words could be weighted more heavily so they are not overshadowed by the more popular yet less influential sentiments.

Regarding the speculative stocks, the results of table 5.2 show the negative and uncertainty sentiments as statistically significant, especially when forecasting beyond same-day price fluctuations. The combination of negativity and uncertainty found here parallels Kahneman's notion of disagreement among investors on the implication of new information (Kahneman, 2003b). The presence of uncertainty creates ambiguity and more room for interpretation from investors. As a result, the regression analysis picks up on this and shows that the presence of uncertainty in press releases is significant to prediction accuracy depending on the timeframe.

In addition, table 5.4 as a whole shows that there is a noteworthy relationship between the number of times each sentiment appears in the Random Forrest model, and the resulting predictive accuracy. The effect sizes here for speculative stocks are much higher than they are for non-speculative stocks. This finding is also in line with the notion that speculative stocks are more prone to being influenced by sentiment than non-speculative stocks (Baker & Wurgler, 2007; Hribar & McInnis, 2012).

The largest effect size in table 5.4 corresponds to the uncertainty sentiment when predicting positive fluctuations in the 48-hour time frame with a value of 0.97. Moreover, the second and third highest effect sizes also appear in the 48-hour time frame when predicting positive price fluctuations, which are 0.93 for negative sentiment and 0.92 for positive sentiment. Again, it is interesting that the largest effect sizes appear most often in positive predictions. While these sentiments do positively contribute to the predictive accuracy, they may be enough to account for broader market forces or other constraining factors.

Returning to answer research questions four and five, the constraining sentiment contributes most to predictive accuracy for non-speculative stocks, while for speculative stocks it

is primarily a combination of the negative and uncertainty sentiments. Therefore, there is a difference in most predictive sentiment, which depends on a stock's level of speculation.

Identifying the Words and Phrases

Through examining the machine learning results, regression analysis and Pearson correlation it has been shown that different levels of speculation and time frames lead to different sentiments being statistically significant and most influential in contributing to predictive accuracy. For non-speculative stocks, the constraining sentiment appears to be the most important, while for speculative stocks it is a mix of the negative and uncertainty sentiments. By isolating the specific sentiments figures 1.1 to 1.6 can be used to provide insight into what words contribute most to predictive accuracy. Furthermore, these words can be traced back to their source and shed light on their context.

Starting with non-speculative stocks, the constraining words used most often were "commitment" and its various conjugations by a large margin, followed by words like "depends," "requirements," "limiting," "restriction," and so on. Examples of sentences from the press releases containing these words are listed here:

1. We are pleased to formalize the **commitment** we have long held to environmental, social and governance issues. (Scotiabank, accomplishment)

2. The flooding has caused many challenges for the people of New Brunswick and through Scotiabank's donation to Canadian Red Cross, we hope some of these challenges will be eased. We recognize each of our customers has been impacted differently by the flooding and we remain **committed** to working with them to help accommodate their individual needs. (Scotiabank, donation)

3. RBC Wealth Management and City National Bank have a **commitment** to diversity and inclusion that dovetails with our work at PowHerful. (Royal Bank of Canada, donation)

4. A.M. Best notes that premium growth **depends** upon the strength of the Canadian and global economy. (Royal Bank of Canada, release of financial strength rating)

5. This project further enhances the bank's **commitment** to support low and moderate income families in our communities, including Veterans who have put their lives on the line to serve our country. (Toronto Dominion Bank, donation)

The trend regarding the word "commitment" is that it corresponds usually with positive events, such as accomplishments or making a donation to a charitable organization. This observation lines up with the results of the Pearson correlation, which found a moderate effect size between the constraining sentiment and predictive accuracy of positive fluctuations. By looking at these sample statements above it would seem that altruistic behaviour and records of accomplishments do positively impact their stock price. However, given that these events do not occur frequently enough, the sentiment generated by these events is drowned out.

This exploration also serves as a reminder that this process is not perfect. The use of the word "depends" corresponds to a boilerplate statement regarding a review of financial strength. Coincidentally, the rating given to RBC during these reviews was positive, which still falls in line with the observed results from the Pearson correlation test.

For speculative stocks, the negative sentiment words that appeared most often were words like "close," "critical," "breaking," "late," "failed," "mislead," etc. Examples of sentences containing negative words are listed here:

1. Shares of Tilray saw double digit gains while Aurora Cannabis **closed** up almost 10%. (Aurora Cannabis, attorney general resigns)

2. InvestorsObserver issues **critical** PriceWatch Alerts for ACB, CHK, MSFT, T, and XSPA. (Aurora Cannabis, third party news release)

3. **Breaking** News: Congress Passes 2018 Farm Bill; Lots of Green Potential to Gain Big in this Uncertain Market (Aurora Cannabis, news)

4. The company's positive statements about the business and its operations were materially false and **misleading** throughout the class period. (Cronos Group, class action lawsuit filed against them)

5. The efforts of Canopy Growth and Canopy Health Innovations to develop a range of patented, insurance coverage eligible cannabis-based medicines took a **critical** step forward with the recent receipt of approval to conduct its first in a planned series of clinical trials. (Canopy Growth Corporation, new initiative)

The sample statements above also illustrate how sentiment analysis at this scale is prone to classification errors. It becomes apparent when reading theses statements that most of the words classified here as negative are not being used in a negative context, with the exception of "misleading." The remedy for this would be to either reclassify those words or remove them altogether from the sentiment dictionary. The two most nefarious words here are "closed" and "critical," as "closed" is a neutral phrase to describe where the price of a stock finished trading that day, and "critical" comes from a third party source which functions as more of an advertisement. The word "breaking," while not strictly negative in this context, does however denote important events, which can impact the price either positively or negatively. In light of these confounding variables, it may be worth dismissing the significance of negative sentiment as a valid conclusion.

Meanwhile, some of the most prevalent uncertainty sentiment words were "risk," "believes," "approximately," "anticipate," "pending," etc. Examples of sentences containing uncertainty are listed here:

1. This study also analyzes the market status, market share, growth rate, future trends, market drivers, opportunities and challenges, **risks** and entry barriers, sales channels, distributors and Porter's Five Forces Analysis. (Aurora Cannabis, research report)

2. Aurora will have an **approximate** 9.14% equity ownership stake in CTT upon conversion of the debenture and holds a warrant which enables Aurora to increase its equity ownership to 42.5%. (Aurora Cannabis, business expansion)

3. Sen. Ron Wyden said of the **pending** law, for too long, the outrageous and outdated ban on growing hemp has hamstrung farmers in Oregon and across the country. (Cronos, bill passed)

4. Additionally, JWC **anticipates** that several of its THC-dominant strains will also become available through the Program. (Canopy Growth Corporation, business expansion)

5. Canopy Growth **believes** that it can add value to the market and enable the development of rigorous testing standards for products, while advancing the understanding of the **risks** and benefits of medical cannabis. (Canopy Growth Corporation, business expansion)

A common theme among the presence of the uncertainty is its co-occurrence with information related to business expansion. By nature, business expansion is wrapped in uncertainty and discussion of it usually contains forward-looking vocabulary. Even in instances where the context did not involve business expansion, the subject was something that could change the way the company does business or to provide investors with additional information. Combined with the findings of the regression analysis and Pearson's correlation, it also makes sense that this uncertainty takes more time to be factored into the predictive accuracy of the stock price fluctuations as it tends to play out in the later timeframes.

Limitations and Future Studies

The choice of only using press releases is perhaps the most restrictive aspect of this study. Given the complexity of the stock market and the numerous factors that influence the fluctuations of prices, it follows that sentiment analysis of press releases should never be able to predict those fluctuations with absolute certainty. There are simply too many other variables that are not accounted for. In an ideal scenario, all published pieces of text would be analyzed, each with their own uniquely constructed sentiment dictionary. This of course is beyond the scope of this study, or any other study realistically. However, by isolating one piece of information and a manageable sample of stocks, conclusions can be drawn about that source of information and its implications on those chosen stocks.

The results are also limited by the fact that the sentiment dictionary used was created initially for analyzing 10-K filings. While this is in the same ballpark as press releases, there were still fairly confounding errors in classification of some sentiments. Therefore one area suited for a future study, would be to create a custom dictionary of financial sentiment to analyze press releases. This should improve predictive accuracy and provide less construed insights.

Conclusion

This study used the Loughran-McDonald dictionary to interpret the financial sentiment of corporate press releases. The results of the sentiment analysis were then used to build a Random Forest machine learning model in order to predict stock price fluctuations. The model used in this study was successful in predicting the stock fluctuations of both speculative and non-speculative stocks. For non-speculative stocks, the model successfully predicted same-day price fluctuations with a 55.95% balanced accuracy. For speculative stocks, the model was successful across all time frames, with a balanced accuracy of 55.38% as its highest score. However, the model struggled to accurately predict positive fluctuations for both the speculative and non-speculative stocks. Therefore, to answer research questions one and two, sentiment analysis of press releases can be used to predict price fluctuations and there is no sizable difference in predictive accuracy between speculative and non-speculative stocks. There is however a difference in accuracy across time frame of the prediction, and the type of fluctuation that the model is trying to predict.

Regarding research question three, there is an optimal time for predictive accuracy. For non-speculative stocks this is the same-day prediction and for speculative stocks, this is the 48-hour time frame. These findings tie in with Fama's EMH and Kahneman's PT. For non-speculative stocks, the companies have a deeper history with more tried and tested sources of information and influential factors. As a result, when news is released for non-speculative stocks there should be less uncertainty surrounding the implications of that information, which therefore leads to more efficient price fluctuations. Speculative stocks however, are more unpredictable and do not have the same extensive history or established factors that non-speculative stocks have, which means new information is less efficiently incorporated into price fluctuations.

Finally, the specific sentiments that contribute most to predictive accuracy were isolated, which addresses research questions four and five. The results of the regression analysis and Pearson's correlations show that the sentiments that contribute most to predictive accuracy differ depending on the stock's level of speculation. For non-speculative stocks, the constraining sentiment was the most important sentiment for predictive accuracy. After tracing the sentiment back to where it would have appeared in the original press release it became apparent that news demonstrating good corporate responsibility could be predictive of positive price fluctuations. For the predictive accuracy of speculative stocks, the uncertainty sentiment was shown to be the most important. Tracing the uncertainty sentiment back to its source further supported this finding, as it often appeared in statements involving business expansion or news that impacts the future of the industry. While negative sentiment for speculative stocks was also shown to be significant, tracing it back to its source showed that evidence of misclassification, which harms its reliability.

In short, this study found that the stock market despite all of its complexity and unpredictability, can be interpreted through the use of sentiment analysis. It is also important to remember that this process of gathering text, interpreting it for sentiment, and making predictions based on it, is still in its infancy. This process, born out of advancements in technology is also made possible by the proliferation of digital communication. The fact that so much information is communicated online provides this method of inquiry with the substance it needs to generate new and interesting findings. While sentiment has likely always been a factor that influences stock prices, only now can it be studied. These findings have always been out of reach, but through the combination of technological advancement and the rise of digital communication they can now be discovered.

Appendicies

Appendix A: Web Scraping Code

```
### Packages ###
library(rvest)
library(stringr)
library(lubridate)

#Enter starting URL.
mw_articles <- read_html("https://www.marketwatch.com/search?q=")

#Define variables/items to be scraped.
URL <- mw_articles %>%
  html_nodes(".searchresult a") %>%
  html_attr("href")

TITLE <- mw_articles %>%
  html_nodes(".searchresult a") %>%
  html_text()

DATE <- mw_articles %>%
  html_nodes(".resultlist span") %>%
  html_text()

datetime_clean <- gsub("\\.","",DATE)

datetime_parse <- parse_date_time(
  datetime_clean, "%I:%M %p %m/%d/%Y"
)
datetime_parse

# Convert all ET (Eastern Time) datetime values to
datetime_convert <- ymd_hms(
  datetime_parse, tz = "US/Eastern"
)

#Create data frame to store data from initial scrape.
PRESS_RELEASES <- data.frame(
  URL=URL, DATE=datetime_convert, TITLE=TITLE
)

dim(PRESS_RELEASES)

#Create loop to cycle through scraped URLs to retrieve the body content.
bodies <- c()
for(i in PRESS_RELEASES$URL){

  PR_COMPLETE <- read_html(i)
  BODY <- PR_COMPLETE %>%
```

```
    html_nodes("#article-body p") %>%
    html_text()
  one_body <- paste(BODY, collapse=" ")
  bodies <- append(bodies, one_body)

}

#Attach scraped body text to data frame.
PRESS_RELEASES$BODY <- bodies

#Remove extra spacing from body text.
clean_text_bodies <- str_squish(
  PRESS_RELEASES$BODY
)

#Create new data frame with squished body text.
PRESS_RELEASES2 <- data.frame(
  URL=URL, DATE=DATE, TITLE=TITLE, BODY=clean_text_bodies
)

dim(PRESS_RELEASES2)

#Create CSV file using the final data frame.
write.csv(PRESS_RELEASES2, "STOCK_MW.csv")
```

Appendix B: Sentiment Analysis Code

```
### Packages ###
library(dplyr)
library(tidyr)
library(stringr)
library(tidytext)
library(ggplot2)

# Load text and explore.
stock.raw <- read.csv("STOCK_FILTERED.csv", stringsAsFactors = FALSE)

# Unnest tokens by BODY.
stock_tokens <- stock.raw %>%
  unnest_tokens(word, BODY, token = "words")

stock_tokens

# Score the TFIDF of BODY by DATE.
date_tf_idf <- stock_tokens %>%
  count(DATE, word) %>%
  filter(!word %in% stop_words$word) %>%
  bind_tf_idf(word, DATE, n) %>%
  arrange(-tf_idf)

# Interpret and visualize financial sentiment.
date_tf_idf %>%
  count(word) %>%
  inner_join(get_sentiments("loughran"), by = "word") %>%
  group_by(sentiment) %>%
  top_n(5, n) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~ sentiment, scales = "free") +
  ylab("frequency of words by sentiment")

# Interpret sentiment by DATE.
date_sentiment_count <- date_tf_idf %>%
  inner_join(get_sentiments("loughran"), by = "word") %>%
  count(sentiment, DATE) %>%
  spread(sentiment, n, fill = 0)

date_sentiment_count

write.csv(date_sentiment_count, "STOCK_DATE_SENTIMENT.csv")
```

Appendix C: Machine Learning Code

```
### Packages ###
library(ggplot2)
library(caret)
library(randomForest)

### Load text and split. ###
data <- read.csv("STOCK_DATE_SENTIMENT.csv", stringsAsFactors = FALSE)
str(data)

# Convert our class label into a factor.
data$same.day <- as.factor(data$same.day)
table(data$same.day)

# 70/30 split for data partition. For replicatability uncomment seed below if desired.
#set.seed(123)
ind <- createDataPartition(data$same.day, times = 1,
                           p = 0.7, list = FALSE)

train <- data[ind,]
test <- data[-ind,]

prop.table(table(train$same.day))
prop.table(table(test$same.day))

# Clean up columns not in use. Comment out one in use.
train$date <- NULL
#train$same.day <- NULL
train$X24hr.change <- NULL
train$X48hr.change <- NULL
train$X72hr.change <- NULL
train$X96hr.change <- NULL

test$date <- NULL
#test$same.day <- NULL
test$X24hr.change <- NULL
test$X48hr.change <- NULL
test$X72hr.change <- NULL
test$X96hr.change <- NULL

### Random Forest ###
# For replicatability uncomment seed below if desired.
#set.seed(456)
rf <- randomForest(same.day~., data = train,
                   importance = TRUE)
print(rf)

# Prediction and Confusion Matrix of training data.
p1 <- predict(rf, train)
```

```
confusionMatrix(p1, train$same.day)

### Prediction and Confusion Matrix of test data. ###
p2 <- predict(rf, test)
confusionMatrix(p2, test$same.day)

# Error rate of Random Forest.
plot(rf)

### Breakdown of variables used in Random Forest model.
# Variable importance.
varImpPlot(rf,
        sort = TRUE,
        main = "Variable Importance")
importance(rf)
varUsed(rf)
```

Appendix D: Regression Analysis Code

```
### Packages ###
library(dplyr)
library(ggplot2)
library(randomForest)

# Import and check data.
data <- read.csv("speculative_average.csv")
str(data)

cols.num <-
c("CONSTRAINING","LITIGIOUS","NEGATIVE.1","POSITIVE.1","UNCERTAINTY")
data[cols.num] <- sapply(data[cols.num],as.numeric)
sapply(data, class)
str(data)

# Create objects for effected.
negative <- data$NEGATIVE
negative <- as.numeric(negative)
positive <- data$POSITIVE
positive <- as.numeric(positive)

# Create objects for effector.
constraining <- data$CONSTRAINING
litigious <- data$LITIGIOUS
negative_sentiment <- data$NEGATIVE.1
positive_sentiment <- data$POSITIVE.1
uncertainty <- data$UNCERTAINTY

# lm regression. Financial Label ~ Sentiments.
summary(lm(negative~constraining+litigious+negative_sentiment+positive_sentiment+uncertaint
y))
```

References

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014, 1-6. doi:10.1155/2014/425731

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. The Journal of Finance, 61(4), 1645-1680. doi:10.1111/j.1540-6261.2006.00885.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. The Journal of Economic Perspectives, 21(2), 129-151. doi:10.1257/jep.21.2.129

Boudt, K., & Petitjean, M. (2014). Intraday liquidity dynamics and news releases around price jumps: Evidence from the DJIA stocks. Journal of Financial Markets, 17(January), 121-149. doi:10.1016/j.finmar.2013.05.004

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324

Chen, H., De, P., Hu, Y., & Hwang, B. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. The Review of Financial Studies, 27(5), 1367-1403. doi:10.1093/rfs/hhu001

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417. doi:10.1111/j.1540-6261.1970.tb00518.x

Fama, E. F. (1991). Efficient capital markets: II. The Journal of Finance, 46(5), 1575-1617. doi:10.1111/j.1540-6261.1991.tb04636.x

García, D. (2013). Sentiment during recessions. *The Journal of Finance, 68*(3), 1267-1300. doi:10.1111/jofi.12027

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

Guo, K., Sun, Y., & Qian, X. (2017). Can investor sentiment be used to predict the stock price? dynamic analysis based on china stock market. Physica A: Statistical Mechanics and its Applications, 469, 390-396. doi:10.1016/j.physa.2016.11.114

Hribar, P., & McInnis, J. (2012). Investor sentiment and analysts' earnings forecast errors. Management Science, 58(2), 293-307. doi:10.1287/mnsc.1110.1356

Kahneman, D. (2003a). Maps of bounded rationality: Psychology for behavioral economics. The American Economic Review, 93(5), 1449-1475. doi:10.1257/000282803322655392

Kahneman, D. (2003b). A psychological perspective on economics. The American Economic Review, 93(2), 162-168. doi:10.1257/000282803321946985

Kaplanski, G., & Levy, H. (2017). Analysts and sentiment: A causality study. Quarterly Review of Economics and Finance, 63, 315-327. doi:10.1016/j.qref.2016.06.002

Kuhn, M., et al. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Li, B., Chan, K. C. C., Ou, C., & Ruifeng, S. (2017). Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Information Systems, 69, 81-92. doi:10.1016/j.is.2016.10.001

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14-23. doi:10.1016/j.knosys.2014.04.022

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. The Journal of Finance, 66(1), 35-65. doi:10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey: Textual Analysis In Accounting And Finance. Journal of Accounting Research, 54(4), 1187-1230. doi:10.1111/1475-679X.12123

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. The Journal of Economic Perspectives, 17(1), 59-82. doi:10.1257/089533003321164958

Maragoudakis, M., & Serpanos, D. (2016). Exploiting financial news and social media opinions for stock market analysis using MCMC bayesian inference. *Computational Economics, 47*(4), 589-622. doi:10.1007/s10614-015-9492-9

Mathotaarachchi, S., Pascoal, T. A., Shin, M., Benedet, A. L., Kang, M. S., Beaudry, T., . . . Rosa-Neto, P., Alzheimer's Disease Neuroimaging Initiative. (2017). Identifying incipient dementia individuals using machine learning and amyloid imaging. Neurobiology of Aging, 59, 80-90. doi:10.1016/j.neurobiolaging.2017.06.027

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in iran. Environmental Monitoring and Assessment, 188(1), 1-27. doi:10.1007/s10661-015-5049-6

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing, 5(2), 221. doi:10.22364/bjmc.2017.5.2.05

Ranco, G., Bordino, I., Bormetti, G., Caldarelli, G., Lillo, F., & Treccani, M. (2016). Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics. PloS One, 11(1), e0146576. doi:10.1371/journal.pone.0146576

Rossi, M. (2015). The efficient market hypothesis and calendar anomalies: A literature review. International Journal of Managerial and Financial Accounting, 7(3/4), 285. doi:10.1504/IJMFA.2015.074905

Seng, J., & Yang, H. (2017). The association between stock price volatility and financial news – a sentiment analysis approach. Kybernetes, 46(8), 1341-1365. doi:10.1108/K-11-2016-0307

Silge, J., & Robinson, D. (2019). *Text Mining with R: A Tidy Approach.* Retrieved from https://www.tidytextmining.com/

Silge, J., & Robinson, D. (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R."_JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), <URL:http://dx.doi.org/10.21105/joss.00037>.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139-1168. doi:10.1111/j.1540-6261.2007.01232.x

Timmermann, A., & Granger, C. W. J. (2004). Efficient market hypothesis and forecasting. International Journal of Forecasting, 20(1), 15-27. doi:10.1016/S0169-2070(03)00012-8

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297-323. doi:10.1007/BF00122574

Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. Expert Systems with Applications, 112, 258-273. doi:10.1016/j.eswa.2018.06.016

Wickham, H. (2019). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.4. https://CRAN.R-project.org/package=rvest

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.1. https://CRAN.R project.org/package=dplyr

Yuz, T. (2018). A Sentiment Analysis Approach to Predicting Stock Returns. *Medium Data Science.* Retrieved from https://medium.com/@tomyuz/a-sentiment-analysis-approach-to-predicting-stock-returns-d5ca8b75a42

Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T. (2018). A novel data-driven stock price trend prediction system. Expert Systems with Applications, 97, 60-69. doi:10.1016/j.eswa.2017.12.026