# Natural Language Analysis of News and its Relation to Market Activity

Connor Neuendorff

A Thesis

in

The John Molson School of Business

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science (Finance) at Concordia University

Montreal, Quebec, Canada

March 2021

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared
By: Connor Neuendorff

Entitled:

Natural Language Analysis of News and its Relation to Market Activity

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Finance)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

________________________ Chair

________________________ Examiner
Dr. Juliane Proelss

________________________ Examiner
Dr. Nilanjan Basu

________________________ Supervisor
Dr. Gregory J. Lypny

Approved by:

________________________ Graduate Program Director
Dr. Nilanjan Basu

____________________ Associate Dean, John Molson School of Business
Dr. Kathleen Boies

March 2021

ABSTRACT

Natural Language Analysis of News and its Relation to Market Activity

Connor Neuendorff

This paper analyzes the impact of sentiment from headlines in the *Wall Street Journal* on earnings surprises and stock returns of US equities. Negative word counts, lexical analyzers, customized dictionaries, and parts of speech analyzers are used together to determine the efficacy of context-specific sentiment analyzers. As headlines do not follow ordinary language rules and positive and negative words have different connotations in a financial context, five metrics are designed to test how different language analysis techniques capture different information. The results indicate that a combination of custom dictionaries, lexical analyzers, and part of speech analyzers captures sentiment relating to earnings surprise more accurately than simple word counts. All of the metrics are significantly related to next day returns but the variability of the prediction is too large to consider them as part of a profitable trading strategy. The results show there is potential for more complex natural language processing techniques for predicting returns.

Acknowledgement

Since accepting me as an undergraduate research assistant, Professor Lypny has been a role model and motivator to relentlessly pursue my interests. His kindness and encouragement will last a lifetime.

Acknowledgement of Research Contributions

I conceived the idea for this research, designed the text analysis metrics as described, and collected and processed all the data necessary to perform the empirical work. Professor Lypny and I exchanged ideas that guided the evolution of the project. I analyzed the results and wrote the manuscript.

Connor Neuendorff, March 2021

# Contents

# 1 List of Tables

# 2 List of Figures

# 3 Introduction

Perspective is everything, as the adage goes. Life is a blur of indeterminate events in which something bad can be seen as good or something good as bad. A professor helping you with your failed midterm can make you more confident than if you'd simply passed the exam and never went to their office hours. A cruel professor can make you feel small no matter how many answers you get right. The professor controls how the event is perceived, if it was good or bad. If a politician makes a controversial statement, the frame which different media sources place it in can polarize it as good or bad affecting viewer's pepreption.

Studying the effects of media on stock prices has increased in popularity with the improvement and simplification of text analysis techniques. Finding patterns in language before the advent of computers would have been a task only for expert linguists with lots of time and funding. But now, even an amateur with a rinky dink processor can do basic word counts on decades of digitized news papers in a matter of hours. This skill has made it easier to analyze qualitative metrics such as recurring topics, sentiment, and writing style.

Paul Tetlock has related the sentiment of news to stock prices by measuring word counts of negative words in news and relating it to many financial performance metrics. He showed, in 2007, that the negative language of a popular news column in the *Wall Street Journal* had an effect on next day stock returns.

This is not in line with an efficient market. It would be one thing for a newspaper's content to be novel and informative, which in a world of second-by-second newswires and high frequency trading shouldn't be the case. But a daily count of negative words predicting a depressed market the next day? This is certainly not an efficient market.

This study investigates further the effects of positive and negative language on stock returns. By using customized language processing techniques and combining different lexicons, sentiment analyzers, and word scores I aim to further demonstrate that perception, defined as qualitative and subjective measures of performance, affects even the most scrutinized and mathematically cold stock price valuations.

# 4 Literature Review

Von Neumann-Morgenstern's axioms of rational behavior have been critical in defining the utility maximizing rational investor (von Neumann 1953) on which central assumptions regarding decision making theory have been founded. However, expecting a market participant to maximize their utility is like expecting a chess player to maximize their chess game (Thaler 2016). They are trying their best but there are too many considerations to factor in and the observed decisions seem irrational.

Modelling heuristics in decision making has made it possible to understand a wider range of phenomenon in asset pricing and to understand how people behave under uncertainty.

## 4.1 Behavioural Science in Finance

The efficient market hypothesis claims that prices fully incorporate all publicly available information and that market participants rationally price this information as soon as it becomes public. Fama's Efficient Market theory (1970) has been central to financial models since its inception and its explanatory power for teaching market concepts cannot be overstated. Yet there are market phenomena that even his theories could not explain. One such instance being the Post Earnings Announcement drift he described as being 'above suspicion' (Fama 1998).

Thaler's analogy of a chess player failing to make all the best moves in a game of chess because they cannot do all the computations is an example of bounded rationality first described by Herbert Simon in Models of Man (1957). This theory describes agents as rational but with limitations. A chess player may try to optimize their chess game, but they have time limits, cognitive limits, and foresight limits. In this light, market participants are not illogical but, simply put, they are unable to compute every permutation of their investment decisions and therefore must rely on heuristics and emotional feedback.

## 4.2 The Role of Emotion in Decision Making

Emotional response plays an important role in decision making for both retail and institutional investors. Bosman et al. (2017) showed that investors are susceptible to language when interpreting financial information. In their experiment, the language of a news article was modified to either have more positive or more negative words without changing the core message of the article. In this way they showed that the framing of news relevant to a company was important and would affect how investors perceived future performance. Investors were more likely to buy stocks that had more positive words and sell stocks that had more negative words even though both articles were delivering the same information.

Tetlock (2007) showed that the sentiment of the *Wall Street Journal's* column "Abreast of the Market" had predictive powers over the next day's market return. Using a dictionary of positive and negative words from the General Inquirer, he showed that the ratio of positive (negative) words was associated with an upward (downward) movement of markets the following day. He used a vector auto regression to determine whether his pessimism variable, based on the General Inquirer dictionary, had predictive capabilities for the Fama French Small Minus Big Factor.

Tetlock (2007) also showed that the inverse is true. Poor market performance, as measured by the Dow Jones, is predictive of negative media sentiment the following day. He regressed the word count of the General Inquirer word classifications of 'Negative', 'Pessimistic', and 'Weak' finding all three to be negatively related to next day Dow Jones returns.

Loughran and McDonald (2011) showed that word dictionaries like the General Inquirer misspecify words when considered in a financial context and built their own word dictionaries that classify financial terms as positive and negative. They built a data set by hand to show that words that may be considered negative in some contexts, such as 'tax' and 'cost' are not negative in a financial context.

They examined the effects of term-frequency inverse document frequency (tf-idf) as a means of controlling for a word's frequency in a text. The tf-idf standardizes the use of a word throughout a document by considering its use in context. The importance of this standardization can be seen in their results. Initially, their custom finance specific negative words had a negative and significant relationship against the abnormal return in the four days after a 10-K filling, while the non-finance specific negative word list was insignificant. However, when the words were standardized, both dictionaries became significant and negative, and had almost an identical association with future returns.

Jiang (2019) showed that the spoken tone of managers during an earnings call, 10-K's, and 10-Q's also has a large and significant effect on aggregated stock market movements. The tone of these managers, as measured by positive and negative word counts, was able to predict downward movements in the aggregate market in subsequent periods. The likely cause of this, as explained by the author, is manager optimism correction. Managers have a more positive tone when their earnings are high and therefore invest more. This sentiment driven investment is later corrected by the market when the earnings are released.

The context in which we are most affected by sentiment is telling of the psychology of investors. Lowenstein et al. (2001) proposed that decisions made under stress are more reliant on emotion than cognitive assessments. Their study showed that the affect experienced during the decision-making process of stressful situations is a greater predictor of how they will proceed in similar situations in the future.

In this psychology review, the authors discussed several experiments in which participants behavior was modified predictably by stress. One example was students tasked with giving a speech for remuneration were asked just prior to the speech if they would reconsider giving the speech. A significantly larger portion of students opted out of the speech after watching a two-minute clip from a scary movie.

They also discussed the effects of self-reinforcing fear leading to panic. Fear leads to increased arousal and this arousal, in turn, leads to increased susceptibility to fear. This feedback loop can cause masses to panic about certain social events, even if there have not been any recent changes to these events. Taken in a financial context, Garcia (2013) showed that stocks are more susceptible to the sentiment of news during recessions.

Again, with the use of positive and negative word counts, he showed that the susceptibility of investors to higher negative word counts in the *New York Times* is significantly amplified during recessions. During non-recession periods there is a positive auto correlation of returns which does not exist during recessions. The effects of pessimism in news during a recession period has a four times larger effect on subsequent returns than during a non-recession period.

There is evidence that we assign more weight to our feelings when we don't have other information available. Baker and Wurgler (2006) showed that companies that are more difficult to value and arbitrage are more susceptible to sentiment. A stock that is difficult to value would be a stock with low capitalization, young, unprofitable, high volatility, non-dividend paying, and distressed firms.

The proxies used for sentiment in their paper are not based on textual analysis but instead closed-end fund discounts, NYSE share turnover, and dividend premium.

Whether it is institutional investors or retail investors that are responsible for sentiment-based trading is not known for sure; however, the evidence indicates that both investors are susceptible to these kinds of heuristics. Barber and Odean (2008) showed that it is both retail and professional investors who are affected by attention grabbing media. They observe trades from discount brokerages and mutual funds and note that when there is a spike in media coverage or large absolute one day return, both mutual funds and retail investors are net buyers.

DeVault (2019) argues that it is in fact institutional investors that are net buyers during high sentiment periods, and it is the retail investors that are the net sellers. He observed that during high sentiment periods institutional investors hold high volatility stocks but the results overall are mixed.

## 4.3 The Effects of News on Stock Returns

The news has a wide range of effects on stocks. Both stock-specific and macroeconomic news have been shown to affect stock prices. Tetlock et al. (2008) says that because most investors don't observe a firm's production activities directly, they have to get the information to make their investment decisions second hand, from the news. He showed that negative firm-specific news is concentrated around earnings announcements, is predictive of lower earnings announcements when it's highly negative, and is also used by analysts as it predicts lower earnings surprises.

Calculating sentiment as the ratio of negative words to the total words of all news stories pertaining to a company mentioned in the *Wall Street Journal* in a 30-day window prior to earnings announcement, he showed a negative and significant relationship between sentiment and unexpected earnings calculated with a seasonal random walk. Using this same sentiment metric, he related sentiment to abnormal return the following day after an earnings announcement.

This shows two things. The first is that the news carries information that is not available through typical financial analysis. News stories carry opinion reflecting qualitative facts that aren't easily extracted from company financials. Secondly, when a company is mentioned in the news, the qualitative information used by analysts to update their forecasts is also absorbed by investors creating an almost immediate price response to non-qualitative information. There is no apparent reversion to the price adjustment following mentions in the *Wall Street Journal*.

Yang et al. (2015) showed that sentiment spikes and consecutive days of same sentiment cause price drift in stock indices. Using a lexical approach for measuring sentiment, they show that when the sentiment of news relating to financial markets is extreme, the SPDR S&P 500 ETF behaves predictably by drifting upwards after good news and downward after bad news for several days after the spike.

Price drifts are indicative of both information updating and slow reactions to information as shown by Chan (2003). He contrasts price movements for news stocks and non news stocks wherein news stocks exhibit a price drift without reversal and no-news stocks eventually reverse. The perceived cause of these price movement patterns is that investors overreact to private information and underreact to public information. Because private information wouldn't appear in newspapers and everything that is published in a newspaper is considered public, the no-news portfolios revert, and the news portfolios continue to drift in the same direction.

Engelberg et al. (2018) showed that an anomaly portfolio based on 97 anomalies has a 50% higher return on corporate news days and six times higher on earnings announcement days. The anomalies, based on a study by McLean and Pontiff (2016), earn these drammatic returns because of biased expectations. Investors are either too optimistic about a stock or too pessimistic which is why these anomalies exist. On news days, the incoming information adjusts these biases and investors revise their expectations causing a premium.

The use of a machine learning algorithms designed to capture the effect of news has had strong results in predicting stock returns. Shumaker et al. (2009) use a combination of bag of words, noun phrases, and named entities to train a machine learning alogrithm to predict returns based on a combination of these factors. They also find that proper nouns are the strongest predictors of returns. They use a support vector machine which is a supervised learning model. This means that it requires training data to find a minimization pattern to classify data. Supervised learning provides challenges with unstructured or novel datasets as it can be very difficult to pre-classify enough data to train the model.

## 4.4 The Effects of Attention

While news has been shown to convey sentiment and inform investors, one of its most impactful features is that it captures the attention of the market. Barber and Odean (2008) proxy investor attention using abnormal change in price, abnormal change in volume, and mention in the news. They find that investors, and specifically individual investors, are more likely to purchase stocks that have attention grabbing features. Because there are so many stocks to chose from, investors are much more likely to chose companies that have attention grabbing features, not entirely unlike advertising in consumer goods.

Gervais et al. (2016) argue a similar point by demonstrating that companies that experience volume shocks become more visible and therefore earn a premium. Stocks that have recently had very poor performance are more likely to have been discarded and subsequently fall off an investors radar. Conversely, volume shocks, regardless of whether the volume shocks are positive or negative, draw attention to a stock causing it to re-enter investors field of vision.

Attention grabbing is not exclusive to individual securities. Yuan (2015) shows that even macroeconomic and broad mentions of the market have an impact on aggregate market returns and individual investor attention. When the Dow Jones Industrial Average, NASDQ Composite Index, or S&P 500 Index make the front page of major newspapers the market reacts predictably behind individuals sell their portfolio holdings more and the market experiences a negative return on the following day.

This occurs for two reasons. The first is due to the disposition effect. Investors sell winners and hold on to losers after a market wide attention-grabbing event. The second is that investors re-balance their portfolios after the market has been brought back to their attention. That is, market events are a signal to investors to do something. Even if it is a financially unsound heuristic like the disposition effect.

Using Google Trends Da et al. (2011) were able to measure when investors were paying attention to a specific security. Google Trends data offers a metric for how many Google searches for a specific term occurred within a time period. They found that a stock with abnormally high search volumes experienced a temporary premium which eventually reversed. This premium was also shown to be most likely caused by individual investors by comparing the search volumes with SEC Rule 11Ac1-5 forms.

Google Trends is an excellent way to look at what people are thinking. In one thoughtful experiment, Preis et al. (2013) were able to capture the effects of panic using Google Trends. By using key words like 'debt', 'housing', and 'inflation' to signal individual financial distress, they were able to predict major market movements based on investor sentiment.

Using the relative change in search terms of key words, they form theoretical long and short portfolios to measure the gains related to investor behavior. The most profitable was a long position in the Dow Jones when the search term 'debt' was the key word. Using 10,000 random simulations and a time frame of 3 weeks to measure the largest change in the search term, their portfolio would have earned a profit of 326%.

## 4.5 Earnings Announcements as Attention Grabbing Events

Earnings announcements are a time when investors will learn if their forecasts were correct and are therefore a time when investors are more likely to monitor their portfolios. They also provide interesting insights into how investors manage their expectations and reactions to actual earnings. Earnings announcement reactions have been shown to behave in illogically predictable ways. Bernard and Thomson (1990) showed that the first three quarters are positively correlated, and the fourth quarter is negatively correlated with the first three. This seems to violate market efficiency as investors do not seem to correctly anticipate this very predictable pattern of post earnings announcement price movements.

Basu (1997) showed that earnings announcement reactions are asymmetric with positive news being incorporated more slowly than negative news. This is because of the Conservatism Principle which states that investors are more reluctant to accept good news than bad news. From an accounting perspective, it is more prudent to always assume the worst so that you can plan accordingly. Basu (1997) showed that the price reactions caused by negative news were also more short-lived than those of positive news.

Because earnings announcements are such a frenzied time for investors, there are surprising heuristics that show how an investors attention can be strained. Della Vigna and Pollet (2009) showed that there is a significantly lower reactions to an earnings announcement on a Friday than any other day of the week. Friday announcements have a lower immediate response and a significantly higher delayed response in comparison to announcements on other days of the week. This shows that something as simple as a Friday afternoon can affect both professional investors and individual investors drastically.

In another display of limited attention, investors have been shown to be susceptible to Kahnman and Tversky's (1974) anchoring effect wherein an individual fails to fully adjust a prediction when presented with an initial value (the anchor). For example, if someone was to place a bid on a house, it has been shown that their bid would be much higher if the asking price was 1 million dollars than if it had been 500 thousand dollars, even if the house was the same in both offers. Context dependent choice (Kahnman & Tversky, 1979) also shows how important decisions can be distorted when given a benchmark value.

Hartzmark et al. (2018) showed that the contrasting effects of other earnings around an announcement will frame the post earnings price change. An announcement that had many negative announcements from other firms in the days preceding its own announcement would have a more positive reaction than if it had no announcements preceding it. And an announcement that had many positive announcements from other companies preceding its own announcement would have a more negative reaction than if it had no announcements preceding it. The authors were able to show that these results were not caused by information transmission and that instead these biased reactions were caused by the context of the reference earnings affecting performance perception.

Investors are also subject to being distracted when many earnings are announced on the same day. Dubbed the 'investor distraction hypothesis', Hirshleifer et al. (2009) show that the post earnings announcement drift is much stronger for firms that have many competing announcements on the same day as their own announcement. Immediate changes in the company's stock price are much smaller however, on busy announcement days as investors scramble to digest all incoming information. Larger firm announcements as well as industry related news are also shown to have a strong impact on investor distraction.

The effects of sentiment on prices during attention grabbing events like earnings announcements have been documented by Mian and Sankaraguruswamy (2012). They show that during high sentiment periods, positive earnings surprises earn higher returns and that during low sentiment periods, negative earnings surprises earn lower returns. Their metric for investor sentiment is the same as in Baker and Wurgler (2006): closed-end fund discounts, NYSE share turnover, and dividend premium. The authors also show that these sentiment-based price effects are temporary and reverse in the long run.

## 4.6 Post Earnings Announcement Drift

The disposition effect (Shefrin & Statman, 1985) plays an important roll in describing the anomalous Post Earnings Announcement Drift. The behavioral heuristic is caused by a combination of mental accounting (Thaler 1985) and prospect theory (Kahneman & Tversky, 1979). The disposition effect is described as the propensity of investors to sell their winning stocks and hang on to their losers.

The cause of this economically unsound behavior is a combination of an innate aversion to loss and a misguided view of how we account for our assets. Say an investor is holding a stock at a loss after the most recent earnings announcement. This investor should sell their asset to be able to reinvest in an asset that is more likely to be successful. Instead the investor holds onto this asset because once they have sold their asset, the mental accounting of having declared a loss is more upsetting than simply holding onto the asset sitting at a loss as it could potentially regain its value.

The role of prospect theory is that investors have asymmetric views of risk depending on whether they are faced with a loss or a gain. Say an investor is holding a stock at a gain after the most recent earnings announcement. In this case, they will sell their asset when they should hold onto it. They sell it because if they are risk averse when faced with gains and their mental accounting tells them that a closed winning position is money in the bank.

These heuristics are not exclusive to individual investors. In his 2006 paper, Frazzini shows that even professional mutual fund managers are susceptible to the disposition effect (Frazzini 2006). He also shows that the disposition effect is at least in part responsible for the post earnings announcement drift. Using brokerage data he is able to see at what price assets were purchased and sold. The capital gains overhang is the difference between the market price and the purchase price, and is the basis for his measurement of investor profit and loss.

When professional fund managers are holding their stocks at a loss after an earnings announcement, they are likely to hold onto these losers, as per the disposition effect. Conversely, if they are holding their assets at a gain, they will sell them rather than hold onto them. This causes a delayed reaction to earnings news as investors do not adjust their portfolios as they should, and this delay slowly corrects itself over the next quarter. Professional investors are subject to irrational thinking which leads to greater financial loss. This shows not only retail investors, or inexperienced investors, are sometimes their own worst enemy.

Weisbrod (2019) builds on the work of Frazinni (2006) by examining the hazard of sale for stocks after their earnings announcement. Inconsistent with an exclusively preference-based model, he finds that the hazard of sale is higher for stocks that are at a large loss. Prospect theory predicts that investors are risk seeking when in a losing position and risk averse when in a gaining position. However, contrary to this preference-based theory, Weisbrod finds that the hazard of sale is an inverse V shape which means investors are more likely to sell when returns are closer to zero indicating they are more likely to speculate when in high loss positions. This again demonstrates how the empirical study of decision making by professionals reveals obvious faults in their decision-making process.

Hirshleifer et al. (2003) showed that the post earnings announcement drift is not caused by individual investors but that they are net buyers and sellers during this time. Using discount brokerage data they are able to see the trading patterns of individuals. They classify individuals into either skilled or un-skilled groups based on past performance and measure how each group trades after an earnings announcement. They find that while both skilled and unskilled investors are net buyers after extreme earnings surprises, their buying patterns are not responsible for the post earnings announcement drift. We therefore cannot explain away the Post Earnings Announcement drift anomaly by blaming individual or amateur investors, and the evidence even suggests that professional investors are not much more clear headed than individuals.

## 4.7 Media and Information Updating

Fairness of information access is very important in maintaining competitive financial markets. If information asymmetries have too large an effect, then only investors with access to a lot of capital or insider knowledge would be able to participate. This means that the diffusion of public information is relevant to both professional and non-professional investors at every level. Merton (1987) discusses the surprising effect of investor recognition on return. While smaller firms generally earn higher returns, this effect is reversed when investor recognition is included in the model. That is, larger firms will earn a higher alpha when investor recognition is accounted for.

Investor recognition is measured as the relative size of the investor base of a security. By comparing the elasticity of expected return with investor recognition, firm size, and firm specific variance Merton argues that an increase in all three properties will lead to lower excess return. His complicated mathematical derivation of this conclusion, based on some of the axioms or rationality previously mentioned, makes intuitive sense, as he demonstrates, if considered from a simple risk aversion perspective.

He shows that more widely known firms earn lower alphas compared to lesser known smaller firms because of 'shadow costs'. Expected returns from factors other than market risk earn smaller, lesser known firms a larger return. The availability of information and its diffusion mediums for larger firms provides investors with information that helps reduce their risks. This type of information is not as easily available for lesser known firms as there are less reports outside of the ones required by market regulations.

Varian (1985) showed that greater dispersion of opinions result in a lower equilibrium market price. Using an Arrow-Debreu model he shows that the equilibrium price is a decreasing function of consumption and that heterogeneous expectations reduce this price. Dispersion of beliefs causes an asset to become riskier and therefore have a lower equilibrium price. This is analogous to Merton's model in which he shows that smaller firms with higher information asymmetries earn higher returns because of increased risk caused by "shadow costs". The more information that is available for a firm, the more investors can make informed decisions at lower risk.

Frieder and Subrahmanyam (2005) showed that individual investors prefer firms with high brand recognition. Using a survey to measure brand recognition they showed that individuals prefer firms with high brand recognition because it reduces information asymmetries and creates a mechanism under which investors can make complicated decisions using simple rules of thumb as outlined in Kahneman and Tversky (1973, 1982). This is consistent with Merton and Varian in that investors will take the tried-and-true asset with lower risk and lower return.

An empirical model to test Merton's theory of investor recognition was built by Fang and Peress (2009) to show that firms with high media coverage have lower returns. By collecting news stories from the largest newspapers in the US and calculating the excess return of high news firms compared to low news firms, they find that high news firms earn almost five percent less than low news firms on an annualized basis. They determine that the primary cause of this effect is the Investor Recognition Hypothesis. By sorting stocks by analyst coverage and individual ownership they show that the effect of media coverage for low recognition firms is much stronger. Firms with high idiosyncratic volatility have a larger response to the media effect as predicted by Merton in 1987.

# 5 Data and Methodology

I build a methodology which enables a more detailed analysis of news as it relates to stocks, or any other financial instrument. My literature review shows that there is extensive interest in relating news to stock performance. With advances in natural language processing, the ability of researchers to analyze millions of news articles, written over decades, and relate their content to firm performance is now not only possible, but depending on the depth of the analysis, easy. The difficulty lies in the details. If you wanted to measure the impact of, for example, a company being mentioned in the news then you would find a digital copy of the news, split up the sentences, and compare each word to the company name creating a tally of the number of occurrences. This information is interesting and useful but it is also only one example of the now seemingly infinite possibilities involved in natural language processing. An example of a more complicated exercise would be training a machine learning algorithm to identify patterns in writing that express fear and anxiety. Another would be developing a linguistics based approach to identify if the author of a text is being deceitful. These more complicated tasks all revolve around around the same data and don't necessarily require any special computing powers. They are examples of the depth and dimension natural language processing has made possible in a financial context. What people say and write, and how they say and write it can, in some instances, be more telling than the financial numbers relating to a firm. Financial models are central to analyzing a firm but, increasingly, the qualitative nature of language in financial documents, news reports, and press briefings are important as well.

## 5.1 News Collection

I analyze the effects of news on abnormal return and earnings surprise, closely following Tetlock et al. (2008) which will, for the sake of brevity, be referred to as Tetlock. Their 2008 study analyzed the effects of the negative words, as identified by the Harvard dictionary of negative words, on both abnormal return and earnings surprise. Using the *Wall Street Journal* as their news source, they find significant and negative relationships exist between their dependent variable: negative word count and their independent variables: abnormal return and earnings surprise.

I recreate this experiment with a few changes. The *Wall Street Journal* is a subscription based newspaper which I am unable to access without a membership. It also requires an expensive license to download news articles in the volume needed for this study. What is available for download is their archive of headlines from https://www.wsj.com/news/archive. The information that is available from this archive has changed drammatically since I started this paper except for the headlines and the date the article was published. The description of the article, body of text, time of publication, and author have changed in their availability and are therefore not included.

To download this data I wrote a web crawler in the C# language which downloaded all of the headlines and the date of publication from January 2000 to end of 2018 which provides 900,442 total news stories. December 2018 was not included because the *Wall Street Journal* changed the structure of their website for that month and I wasn't able to download it.

As this data was scraped from the original source, it is unstructured. That is, there are no labels of any kind in this very long list of headlines and dates. What this means is, to identify if a company is mentioned in a headline, I had to first identify how a company is represented normally in a headline. For example, Abbot Laboratories is not mentioned as "Abbot Laboratories" in a WSJ headline but is instead mentioned as "Abbot". Advanced Micro Devices is mentioned as "AMD". Southwest Airlines can be mentioned as both "Southwest Airlines" or "Southwest". Creating a list of companies exactly as they are mentioned with all their variations would require reading every headline and is beyond the scope of this study. I instead developed a list which minimized false positives.

Starting with a list of every company in CRSP, which totals 5400, I develop a proceedure to adjust the names so as to remove companies which cannot be uniquely identified in a headline. I manually go through the list of companies and adjust the name to what I think it will appear as in the headline. Certain companies are discarded immediately because they won't be identifiable without returning headlines unrelated to the specific company. Exactly which companies are uniquely identifiable and which are not is dependent on the dataset being used. I manually removed some pre and post fixes and adjusted the company names incrementally. It is easily tell that Walmart Inc is not mentioned with the post-fix "inc" however it is less intuitive that Square Inc would be mentioned with it's post-fix. After each adjustment to a company name the adjusted name is run through the dataset to see what headlines are returned. If they matched headlines are correct and there are little to no false positives then the company is included in the final list.

The final number of companies was 523 with some companies having multiple variations of their mentioned names. To be able to include a company in this list it had to be mentioned at least once in the 18 year sample. This list of companies included any company, of any size that is in the CRSP database. The price data for each security was retrieved from CRSP. The earnings data is retrieved from I/B/E/S. The quarterly and financial data used for control variables is retrieved from Compustat. The GVKEY's, PERMNO's, and CUSIP's are all used to combined the data. If a company is not available in CRSP, Compustat, and I/B/E/S it is not included.

| | |
|---|---|
| Number of WSJ headllines Jan 2000 - Nov 2018 | 900,442 |
| Number of news days Jan 2000 - Nov 2018 | 6,509 |
| Starting Number of companies | 5400 |
| Number of companies found in WSJ | 523 |
| Number of headlines which mention a company | 82,673 |

## 5.2 Sentiment Analyzers

Natural language processing techniques used to process a headline vary considerably based on intended outcome and context. I compare five sentiment analysis techniques, each designed to capture emotional content in the hopes of relating emotion to financial decision making patterns. Three of these five I built in the hopes of capturing sentiment that is specific to a financial context and specific to headlines. As both of these contexts have different attributes it makes sense to design an algorithm for this context.

The first sentiment analyzer is the same one that is used in Tetlock. They use a negative word count which uses the *General Inquirer's Negative* word list as its basis. They do not include positive words due to the weakness of the results. To count the negative words in a news story, they combine all of the company-specific headlines of a day into one headline. They then count the number of negative words and compare it to the total number of words. This variable is called Neg. Neg is then standardized by subtracting the prior years average and standard deviation. This standardized variable adjusts for potentially non-stationary mentions in the news. neg is therefore considered a stationary variable. This standardization is used for all the sentiment metrics in this paper. By using the total words per day instead of counting the sentiment on a per story per day basis there is potential that a story which carries a lot of sentiment can be drowned out by the rest of the day's stories. However, I don't expect one news headline to alter the index's return. Instead, what I'm looking to capture is pronounced and ubiquitous sentiment that permeates most of a days headlines. I replicate this metric and refer to it as Tetlock for the remainder of this paper.

$$Neg = \frac{No.\ of\ negative\ words\ relating\ to\ a\ company}{No.\ of\ total\ words\ per\ day\ relating\ to\ a\ company} \tag{1}$$

$$Tetlock = \frac{Neg - \mu_{Neg}}{\sigma_{Neg}} \tag{2}$$

The second sentiment analyzer uses a more advanced lexical analyzer with built in positive and negative dictionaries. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a lexical analyzer designed to determine sentiment of online reviews, twitter sentiment, and other unstructured language. As these types of text do not always follow proper grammar rules, the lexical rules which evaluate them must be flexible. This is useful when analyzing headlines as they also do not generally follow grammar rules.

The VADER package in Python provides a sentiment score normalized between -1 and +1. It allows for more complicated analysis of a sentence by incorporating the parts of speech (POS) in a sentence instead of just using word counts. It can interpret negation, degree modifiers and POS tagging. The sentence "Exxon did not suffer catastrophic loss" would return as negative under a word count scheme but would return positive with a lexical analyzer like VADER. The VADER sentiment score is calculated for each headline on a particular day and that score is then summed for each day. This VADER is then standardized by subtracting the prior year's average and dividing the result by the prior year's standard deviation resulting in Vader.

$$Vader = \frac{\Sigma\ VADER_{daily} - \mu_{VADER}}{\sigma_{VADER}} \tag{3}$$

Headlines in a financial context have an observable pattern involving direction. In ordinary language, up, down, fall, slide do not carry much connotation with respect to sentiment. In finance however, they are often the defining characteristic of a headline. "Stocks fell after the announcement from the Fed", "Earnings rise after quarterly revision", and "Disney's Stock Price Dropped After Shareholder Meeting" are all examples of headlines that would have returned neutral sentiment scores without the addition of directional words to the lexicon.

The third sentiment analyzer uses directional words in combination with the VADER sentiment analyzer. To add sensitivity to direction, 39 positive and 28 negative words are added to the lexicon. The VADER library in python allows the user to add, remove, or alter the sentiment scores in the lexicon. While the semantic analyzer remains unaltered, a user can change the lexicon enabling a customized dictionary for differing use cases. These words include “rise”, “fall”, “soar”, “drop”, “grow”, “dove” along with the different grammatical tenses and plurality for a total of 67 words representing direction. This variable is then standardized to be stationary over time.

$$Vader_{Directional} = \frac{\Sigma\ VADER_{daily_{Directional}} - \mu_{VADER_{Directional}}}{\sigma_{VADER_{Directional}}} \tag{4}$$

The fourth sentiment analyzer uses the parts of speach (POS) of a headline to focus on the root word and direct object. Using the Python library spaCy, I break the headlines up into their parts of speech and tokenize the dependencies, as explained below. Using existing libraries of sentence structure and dependencies, spaCy is able to identify the different roles of words in a sentence. It is able to identify the different POS and their constituents. As headlines do not follow ordinary sentence structure, this technique allows the sentiment analyzer to focus directly on the key words in a headline.

Parts of speech (POS) define the syntactic function of a word. For example, in the sentence “H&R Block Names Jeffrey Jones President, CEO” the lexical analyzers cannot distinguish the parts of speech for “H&R Block”. It is unclear for the Vader analyzer if “Block” is a proper noun or adjective. In fact, Vader does assume, incorrectly, that “Block” is a verb instead of a proper noun. While natural language processing is not a solved issue in computer science there are several ways to disambiguate POS and correctly interpret language. One way to interpret this headline would be to look at the capital letters. Ordinarily proper nouns are capitalized and in this case, if it had been a sentence and not a headline, this method would have disambiguate the phrase. However, because it is a headline, all the words are capitalized.

A much more complicated but robust way of disambiguating POS is to vectorize a corpus of text and determine the probabilistic context of words. This is done by converting each word into an n-dimensional vector and then use a minimization technique (usually a neural network) to determine which word vectors are most closely related. This allows for probabilities to be calculated to determine what the likelihood is that a word, in a given context matches a particular part of speech; in this context the word “Block”. If in a corpus the word “Block” appears more often in the context of “H&R Block” than it does in an example context of “To Block” then the POS tagger would categorize it as a verb. SpaCy is a pre-trained model which has been built using a neural network to vectorize words based on the context of a training corpus. Because of this, it is able to disambiguate POS. Ideally the POS tagger would be trained on a corpus of *Wall Street Journal* articles which would maximize the accuracy of the tagger; however, that is beyond the scope of this paper as it would require significantly more data than is available.

Instead, I combine POS analysis with the VADER lexical analyzer to caputure another dimension of the headline’s sentiment. The first step is to parse the sentence into word tokens. The dependencies of these tokens are then labeled by spaCy. Example labels are the root word, noun subject, direct object, preposition, etc. I focus on the root word and the direct object of the headline to determine the sentiment of the entire headline. The root connects the relevant noun phrase to the rest of the text and therefore carries a lot of meaning in a headline. It is usually a verb. In the sentence “Chevron missed big gains”, the root word is “missed” and the direct object is “gains”.

By focusing on the root word and the direct object only, we narrow the focus of the sentiment analyzer. Once the root word and direct object are identified, VADER is used to identify the sentiment of each and the resulting score is summed. If the score is greater than 1 then it becomes 1, and if it is less than -1 it becomes -1. Finally, a grammatical rule is added which states that if the the sentiment of the root word is not zero and disagrees with the non-zero sentiment of the direct object, an adjustment is made. If the non-zero root is negative and the non-zero direct object is positive the direct object is multiplied by -1. If the non-zero root is positive and the direct-object is negative the direct object is multiplied by -1 .

This is done to account for the focus of the headline. For example, the headline “Chevron missed big gains”. The root is “missed” which is negative, and the direct object is “gains” which is positive. Without the above mentioned adjustment the result would be neutral. However, with the adjustment the result is correctly negative. Another good example of the success of this adjustment is in the headline “New Novartis Drug may Treat Cancer”. With the positive root “treat” and the negative direct object “cancer”. I developed this rule based on my perception of which words carry the most emotional weight in a headline. There is not a rule in linguistics of which I am aware that explains this relationship or that certifies it as a logical interpretation of POS. Through my analysis of a small sample of headlines this rule showed to be useful in disambiguating the sentiment of a headline within a financial context. My intention was to focus the sentiment analyzer on words I thought would affect the reader the most.

$$x \mapsto sign(root) = \begin{cases} 1, & \text{if } x > 1 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < -1 \end{cases}$$

$$y \mapsto sign(dobj) = \begin{cases} 1, & \text{if } y > 1 \\ 0, & \text{if } y = 0 \\ -1, & \text{if } y < -1 \end{cases}$$

$$(x \neq 0, y \neq 0) \bigwedge (x = y \rightarrow Space = root + dobj) \bigwedge (x \neq y \rightarrow Space = root + (-1 \cdot dobj)) \tag{5}$$

$$SPACE_{daily} = \begin{cases} 1, & \text{if } Space > 1 \\ -1, & \text{if } Space < -1 \\ Space, & \text{if } -1 < Space < 1 \end{cases}$$

where root is the central term of the headline determined using spaCy and the dobj is the direct object.

$$Space = \frac{\Sigma\, SPACE_{daily} - \mu_{SPACE}}{\sigma_{SPACE}} \tag{6}$$

Finally, the last sentiment analysis variable I use combines the use of directional words with VADER and the POS analysis of Space. This variable, SpaceVader, includes the 67 directional words included for the calculation of equation (4), and uses POS tagging for the root word and the direct object for equation (6) to provide a financial context to the headline sentiment analyzer which focuses on the root and direct object.

$$SpaceVader = \frac{\Sigma\, SPACEVADER_{daily} - \mu_{SPACEVADER}}{\sigma_{SPACEVADER}} \tag{7}$$

## 5.3 Comparison of Sentiment Metrics

In this section I analyze the different headlines that each of the five sentiment metrics capture. This illustrates how the varying methodologies for calculating sentiment result in a diverse set of results. Firstly, I have compared each sentiment metric by building a relationship co-occurrence matrix. Table 1 shows the number opf stories captured by both metrics and Table 2 the percentage captured. The value of each element in this matrix is calculated by summing the number of headlines in which both sentiment metrics calculate a non-zero value for that headlines sentiment, and dividing this result by the number of stories captured by the sentiment metric on the horizontal axis. With this methodology, I can measure which percentage of one metric is captured by the other, and vice versa, unlike with a correlation matrix. Each element of the matrix can be evaluated as the percentage of metric in the rows occurring at the same time as the metric in the header.

We can see from Matrix A that Vader returns the most non-zero sentiment scores for news, almost 50% more. This is most likely due to the fact that it incorporates both positive and negative words, unlike Tetlock which only considers negative words. We can see a strong distinction on the sentiment metrics with regard to the number of stories they return. Tetlock is the only one which considers negative exclusively and has 10,000 results. Vader and $Vader_{Directional}$ are in a close range of 13,000 to 14,000 and both consider positive and sentiment words, using almost identical language processing techniques, excluding the 67 additional words in $Vader_{Directional}$. Finally, we see Space and SpaceVader which use different language processing techniques than the previous methods and return a much smaller number of stories than their Vader only counter parts.

We can also see which sentiment metrics are most closely related by looking at Matrix B. Both Vader and $Vader_{Directional}$ have very high co-occurrence values implying they capture all of the stories from Tetlock, Space, and SpaceVader. Some of the metrics do not capture any stories not captured by other, broader, metrics but instead narrow down the result set. If we look at the values for SpaceVader and Vader in Matrix B we can see that SpaceVader only accounts for 74% of Vaders results while Vader accounts for 99% of SpaceVader's results. This shows that SpaceVader does not capture any stories that Vader does not. However, it appears to narrow the result set, reducing the noise or increasing precision as compared to Vader. A lexical analyzer like Vader will cast a wide net but the addition of a POS analyzer will filter out some of the invalid results.

$$
\begin{array}{c c}
\begin{array}{ccccc} Tetlock & Vader & Vader_{Directional} & Space & SpaceVader \end{array} & \\
\begin{bmatrix}
10022 & 9821 & 9420 & 7298 & 8161 \\
9821 & 14411 & 12947 & 9019 & 10685 \\
9420 & 12947 & 13013 & 8998 & 10684 \\
7298 & 9019 & 8998 & 9032 & 8990 \\
8161 & 10685 & 10684 & 8990 & 10730
\end{bmatrix} & \begin{array}{c} Tetlock \\ Vader \\ Vader_{Directional} \\ Space \\ SpaceVader \end{array}
\end{array}
$$

Table 1: Co-Occurrence Matrix A

$$
\begin{array}{c c}
\begin{array}{ccccc} Tetlock & Vader & Vader_{Directional} & Space & SpaceVader \end{array} & \\
\begin{bmatrix}
1 & 0.68 & 0.72 & 0.80 & 0.76 \\
0.97 & 1 & 0.99 & 0.99 & 0.99 \\
0.93 & 0.89 & 1 & 0.99 & 0.99 \\
0.72 & 0.62 & 0.69 & 1 & 0.83 \\
0.81 & 0.74 & 0.82 & 0.99 & 1
\end{bmatrix} & \begin{array}{c} Tetlock \\ Vader \\ Vader_{Directional} \\ Space \\ SpaceVader \end{array}
\end{array}
$$

Table 2: Co-Occurrence Matrix B

$$
\begin{array}{c}
\begin{array}{ccccc}
Tetlock & Vader & Vader_{Directional} & Space & SpaceVader
\end{array} \\
\left[\begin{array}{ccccc}
1 & -0.06 & -0.09 & -0.05 & -0.04 \\
-0.06 & 1 & 0.22 & 0.13 & 0.11 \\
-0.09 & 0.22 & 1 & 0.29 & 0.31 \\
-0.05 & 0.13 & 0.29 & 1 & 0.75 \\
-0.04 & 0.11 & 0.31 & 0.75 & 1
\end{array}\right]
\begin{array}{c}
Tetlock \\ Vader \\ Vader_{Directional} \\ Space \\ SpaceVader
\end{array}
\end{array}
$$

Table 3: Correlation Matrix Between Sentiment Metrics

Some of the variables are highly correlated as shown in Table 3. Space and SpaceVader have a Pearson correlation coefficient of 75%. Vader and $Vader_{Directional}$ have a Pearson coefficient of 31%. These relationships should be expected as the variations between them are small. Space and SpaceVader are only differenced by the lexicons they use and the same is true with Vader and $Vader_{Directional}$. There is very little correlation with the Tetlock variable however and all the other variables are slightly negatively correlated with it. This is possibly caused by the Tetlock variable only capturing negative sent.

The sentiment scores for all five metrics can be see in Figure 1. Several extreme values had to be removed from Vader and $Vader_{Directional}$ to be able to graphically represent their distributions. We can see from these box plots that the first and third quartiles, as well as the median, are zero for all five metrics. The majority of the values for all the metrics are zero. That is because only a small portion of the news surrounding both earnings announcements and ordinary news days will be neutral in nature. Most news, as compared to its own historical scores, will be neutral. However, this figure also illustrates that there could potentially be a lot of high sentiment news that is being returned as neutral due to limitations of the sentiment analyzers to detect financially significant news from headlines. The data also does not appear to be evenly distributed. The direction of the skew changes from metric to metric and is indicative of their differing abilities to capture sentiment.

Below are comparisons for each of the sentiment metrics. For each of the five metrics there is one positive, one negative, and one neutral headline score as per that sentiment metrics result. Several of the same headlines are used for the same metrics to illustrate how each metric returns different results. One example of a headline that returns a myriad of sentiment polarities is “British American Tobacco First-Quarter Sales Fall 12%”. As the reader, we know that this is not

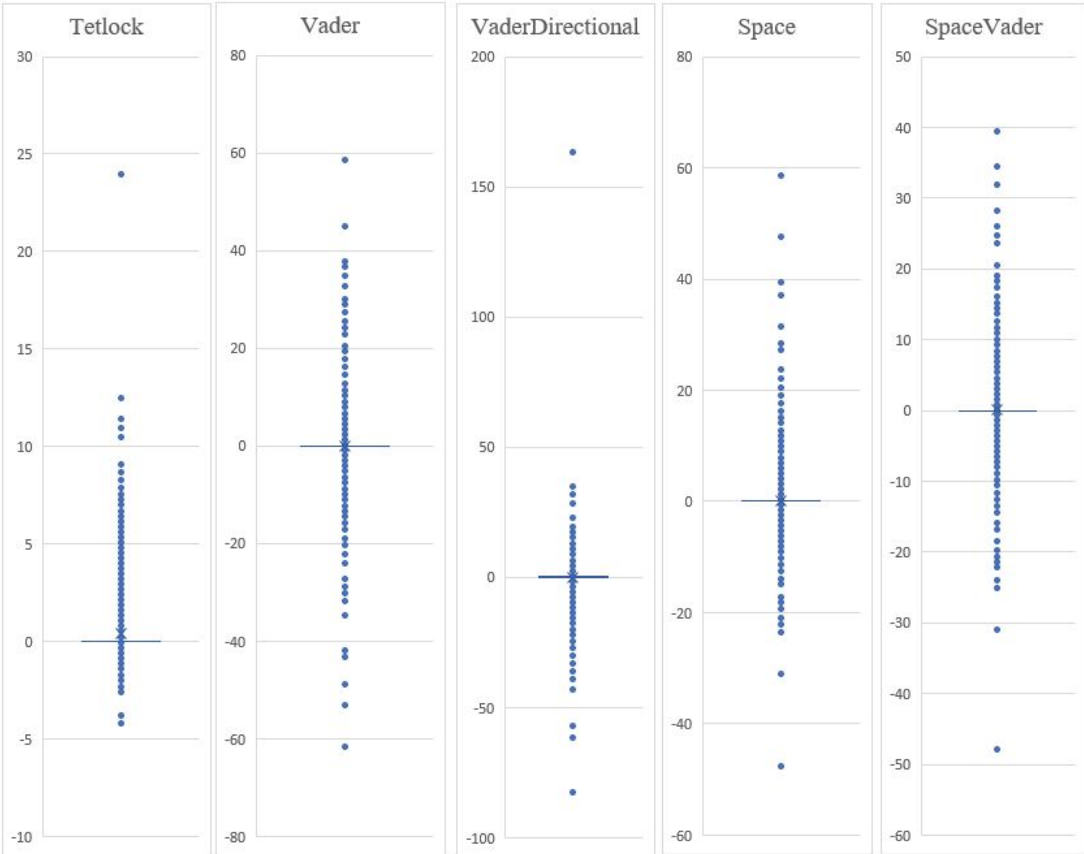


Figure 1: Box Plot: Distribution of Sentiment Scores for 5 Metrics

good news. However, of the five sentiment metrics only two return a negative score.

Tetlock returns this headline as neutral because there aren't any of these words in the General Inquirer negative word dictionary. Vader also returns this score as neutral because it's dictionary hasn't been altered to include directional words in a positive or negative context. The $Vader_{Directional}$ metric however, does return a negative score because it has been trained to associate falling, tumbling, or decreasing as negative. Similarly, Space does not return a negative score but SpaceVader does.

One more example of how these metrics differ can be seen with the headline "H&R Block Names Jeffrey Jones President, CEO". For Tetlock, Vader, and $Vader_{Directional}$ this headline returns a negative score. This is simply because of the word "block" within "H&R Block". This word is in both the General Inquirer negative word dictionary and in Vader's original lexicon as a negative word. In this instance these metrics fail at identifying the parts of speech correctly. It isn't possible for these three metrics to distinguish between "H&R Block" where "H&R" is the Proper Noun and "Block" is the verb and "H&R Block" as a proper noun entirely.

| | |
|---|---|
| | Tetlock |
| Positive | Korea, Israel May Get MSCI Change |
| Neutral | British American Tobacco First-Quarter Sales Fall 12% |
| Negative | H&R Block Names Jeffrey Jones President, CEO |
| | Vader |
| Positive | Outgoing Spirit Airlines Chairman in Talks to Buy Frontier Airlines |
| Neutral | British American Tobacco First-Quarter Sales Fall 12% |
| Negative | H&R Block Names Jeffrey Jones President, CEO |
| | $Vader_{Directional}$ |
| Positive | Magna Profit Climbs 30%, Raises Dividend |
| Negative | British American Tobacco First-Quarter Sales Fall 12% |
| Neutral | Intel Scores Speed Breakthrough |
| | Space |
| Positive | International Paper Posts Profit |
| Neutral | British American Tobacco First-Quarter Sales Fall 12% |
| Negative | Sinopec Daylight Executive Faces Insider-Trading Charges |
| | SpaceVader |
| Positive | Chevron Net Falls; Unocal Profit Rises |
| Negative | British American Tobacco First-Quarter Sales Fall 12% |
| Neutral | Pfizer Bone Drug Gets Limited Nod |

Table 4: Example Positive, Neutral, and Negative Headlines. The metrics have varying results for the same headline. This table shows both the weakness and strength of the metrics as some capture sentiment correctly and some incorrectly.

# 6 Results

This section compares the five sentiment metrics in their ability to predict both abnormal return and earnings surprise. The results from Tetlock et al. 2008 show that headlines have an association with investor behavior. I utilize the methodology from their paper but modify the variables to create context specific algorithms designed to work on headlines. Headlines from the WSJ should not, according to efficient market theory, have any effect on investors so long as the headlines do not hold new information that wouldn't be available to investors through sophisticated news wires. I assume that it is rare that a story will hold new information regarding a company's financial performance. Stories in the WSJ, like most stories in the news, provide insight and analysis into the raw, second by second news releases offered by other financial news services such as reuters news wires.

That is to say, the information contained in the articles analyzed by this paper is not new. We can expect sophisticated investors to be monitoring more timely news sources. That means that these articles shape opinion but do not offer new information. They simply frame and explain the circumstances and provide additional information or historical perspectives. For investors who do not follow more timely sources, these articles offer stale information. Presumably this type of investor incorrectly interprets this information as breaking news and might even perceive these articles as an informational advantage (especially so if they are paying the subscription fees). For investors who do follow more timely sources, such as newswires, should the sentiment metrics be significant, it would be analogous to the results found in Bosman et al. (2017) that showed how language used has a significant effect on how investors perceive information. My hypothesis is that positive sentiment headlines will have a positive effect on abnormal returns and earnings surprise, and negative sentiment headlines will have a negative effect on abnormal returns and earnings surprise.

## 6.1 Earnings Announcements

There is more news around an earnings announcement for a company than there is normally (Tetlock et al., 2008). It is a time when investors are most focused on their portfolios and when investors are more likely to adjust their holdings. I analyze the effects of the news sentiment on the earnings surprise and the abnormal return immediately following the announcement. These two tests will show how sentiment is associated with investor expectations, as measured by earnings surprise, and their subsequent reactions, as measured by abnormal return. Following

Tetlock's parameters closely, I calculate standardized unexpected earnings following Bernard and Thomas (1989) who use a seasonal random walk with trend. That is, there is a serial correlation between each quarter and to determine the seasonal random walk, the quarter from the same quarter last year must be used to detrend the same quarter in the present year:

$$UE = UE_t = E_t - E_{t-4} \tag{8}$$

$$SUE_t = \frac{UE_t - \mu_{UE_t}}{\sigma_{UE_t}} \tag{9}$$

where $E_t$ is the earnings in quarter t, and the seasonal random walk is accounted for using the mean ($\mu$) and standard deviation ($\sigma$) of the firm's previous 20 quarters. As in Tetlock, each firm must have earnings data for the past 10 quarters and if a firm has less than four years of earnings data I assume a non-zero trend. Unlike in Tetlock, however, I do not winsorize the earnings announcement at 99% so as to measure the extremes of all my variables.

For the sentiment metrics to capture any effect prior to the earnings announcement I standardize them for a 28-day window as is done in Tetlock. That is, the five equations mentioned above are summed for a window of [-30, -3] instead of on a daily basis and the mean and standard deviations used to standardize them are also calculated using this time frame. This 28-day window allows for a measure of most of the news that will be relevant to the company's upcoming earnings announcement. The three-day window right before the announcement allows for any news published at t = -3 to be fully absorbed by the market.

To ensure that the analysis of headlines is as similar as possible to Tetlock's analysis of entire stories, I include his control variables exactly as stated. These are lagged earnings, size, book-to-market ration, trading volume, three measures for stock returns, and analyst forecast dispersion. Lagged earnings is measured as the firms previous SUE. Firm size is calculated as ($Log(MarketEquity)$). Book-To-Market is calculated as ($Log(Book/Market)$) from the previous calendar year as in Fama French (1992). The trading volume is calculated as the log of annual shares traded divided by shares outstanding (Log(Share Turnover)) from the previous calendar year.

I calculate abnormal return as the excess of the S&P 500. There are three return periods included as control variables. As in Tetlock, I use two short term abnormal returns as control: $CAR_{-30,-3}$ and $CAR_{-2,-2}$. These measures capture any recent shocks to the stocks return. I also use $CAR_{-252,31}$ to capture the momentum return as per Jegadeesh and Titman (1993). I also include a control variable for Forecast Dispersion measured as the standard deviation of analyst forecasts for the most recent quarter divided by the standard deviation of the earnings volatility of the 10 most recent quarters. All of the abnormal return data is calculated from the CRSP data sets and all of the earnings data from I/B/E/S.

SUE is already standardized but will still have some correlation with the calendar quarter of its announcement as shown in Petersen (2007). Therefore, I use a a pooled ordinary least squares regression and standard errors clustered by calendar quarters to calculate the effects of my sentiment metrics on SUE, as per Tetlock. Table 5 reports the results of my five sentiment metrics calculated using only the headlines from the WSJ.

Table 5 displays the results of the coefficients from all 5 independent variables. We can see that almost all of the control variables are highly significant with the exception of the two short term abnormal returns. The $R^2$ is very high. The majority of this high value is caused by the lagged surprise control variable. This powerful effect of lagged surprise is not consistent with Tetlocks results. A test to determine if this large $R^2$ was the result of multi-colinearity was conducted by calculating the variance inflation factor and no significant colinearity were found. The large impact of lagged surprise merits further investigation but for the purpose of this paper, I am satisfied the cause is purely economic and not a result of a fault in the model. We can see that only the Space sentiment metric is useful in identifying the effects of news headlines on earnings surprise. As the Tetlock variable is not significant, we can say that a more targeted approach to interpreting headline sentiment is a viable solution to sentiment analysis.

By focusing on the most important words in the headline and their relationship with each other, we find a way to classify sentiment in a way that a simple word count cannot. While the effects are small and the significance weak, this indicates that it's possible, with further study of the language of headlines, to identify more adept solutions to quantifying sentiment. This also shows that the sentiment metric is viable for both positive and negative sentiment, unlike a negative word count which is only able to capture negative sentiment. We can see that both positive and negative sentiment affect earnings surprise.

Lastly, we see that the relationship exists for Space and not for the ordinary word count Tetlock. While it is possible that the qualitative content of the headlines is the cause of the relationship, it can be seen here that this content is only perceived by the more advanced language parser. Whether the SUE captures

Table 5: This table compares five separate sentiment metrics against the Standardized Earnings Surprise (SUE) using an ordinary least squares regression. The sentiment metrics are calculated as the sentiment score in the [-30, -3] window standardized by the previous years average and standard deviation for the headlines of the *Wall Street Journal*. All regressions include lagged regression, forecast dispersion, firm size, book-to-market, trading volume, $CAR_{-252,31}$, $CAR_{-30,3}$, and $CAR_{-2,-2}$ control variables. The robust t-statistics are in parentheses.

| Parameter | Tetlock | Vader | $Vader_{Directional}$ | Space | SpaceVader |
|---|---|---|---|---|---|
| Sentiment Mentric | -0.00915 | 0.0074 | 0.027 | 0.0384* | 0.0324 |
| | (-0.36) | (0.44) | 0.75 | (1.67) | (1.40) |
| $SUE_{lag}$ | 0.8092*** | 0.8093*** | 0.8093*** | 0.8092*** | 0.8093*** |
| | (71.37) | (71.40) | (71.52) | (71.44) | (71.50) |
| Forcast Dispersion | 1.9415*** | 1.9419*** | 1.9406*** | 1.0426*** | 1.9439*** |
| | (10.42) | (10.41) | (10.42) | (10.44) | (10.44) |
| Firm Size | 0.8039*** | 0.80392*** | 0.80396*** | 0.8039*** | 0.8038*** |
| | (11.05) | (11) | (10.99) | (11) | (10.99) |
| Book-to-Market | -125.6481*** | -125.6*** | -125.5937*** | -125.2396*** | -124.989*** |
| | (-7.68) | (-7.67) | (-7.68) | (-7.64) | (-7.64) |
| TradingVolume | 2.0873*** | 2.0875*** | 2.0848*** | 2.0891*** | 2.0856*** |
| | (5.89) | (5.88) | (5.87) | (5.89) | (5.88) |
| $CAR_{-252,-31}$ | -0.2049** | -0.2052** | -0.2060** | -0.20889** | -0.2102** |
| | (-2.24) | (-2.24) | (-2.24) | (-2.28) | (-2.30) |
| $CAR_{-30,-3}$ | -0.2745 | -0.2772 | -0.2821 | -0.2934 | -0.2957 |
| | (-0.94) | (-0.94) | (-0.96) | (-0.97) | (-0.98) |
| $CAR_{-2,-2}$ | 0.0584 | 0.0469 | -0.2821 | -0.2940 | 0.7044 |
| | (0.05) | (0.04) | (-0.96) | -0.97 | (0.06) |
| Observations | 14008 | 14008 | 14008 | 14008 | 14008 |
| Clusters | 76 | 76 | 76 | 76 | 76 |
| $Adjusted R^2$ | 0.7952 | 0.7952 | 0.7953 | 0.7953 | 0.7953 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

surprise caused by sentiment or a calculated adjustment based on factors such as earnings or quantifiable data, we can still see the effect of headlines on analysts expectations.

Because the earnings estimates are always compiled by experienced analysts who are trained to not be biased by frivolous language, we can expect any reaction to be small. However, investors are not required to be experienced and immune to sensational headlines and should therefore be much more susceptible to language. We can see the relationsiup of language on investors by comparing the sentiment metrics on the abnormal return following the earnings announcement. Table 6 shows the relationships of all five sentiment metrics on the abnormal return 1 day after the announcement.

The results from Table 6 show that both the Tetlock sentiment metric and SpaceVader sentiment metric are significant at the 1% and 5% levels respectively. This shows that the sentiment of the headlines in the 28 days prior to the earnings announcement have a significant impact on the abnormal return after the announcement. Interestingly the coefficient for the Tetlock variable is positive implying that the more negative the news, the more positive the return after the announcement. It is possible that this indicates a correction from the market where very negative news causes an overly pessimistic view of the companies performance that corrects after the announcement is made. We can also see that the only control variable that is significant is firm size. This could be because firm size is directly related to it's investor base and the different types of investors cause a high degree of variation.

SpaceVader is significant and positive showing that it is able to predict both positive and negative abnormal reactions to news headline sentiment. It is however much weaker than the Tetlock variable. This shows that a simple negative word count outperforms a more context specific sentiment analysis tool. It is possible that the reason for this is that negative headlines carry more weight for investors around earnings announcements and therefore a sentiment metric that focuses only on negative words is more accurate.

## 6.2 Returns in Story Event Time

I now focus on the effects of sentiment on the abnormal returns following a news story. The effects of sentiment on abnormal return around earnings announcement are significant but they are a special case. Earnings announcements are a time of great focus for investors. I therefore measure the impact of news

Table 6: This table compares five separate sentiment metrics against the Abnormal Return (AR) 1 day after the earnings announcement. The independent variable is the abnormal return and is calculated as the excess return of the S&P 500. The regression is an ordinary least squares regression clustered around calendar quarters. The sentiment metrics are calculated as the sentiment score in the [-30, -3] window standardized by the previous years average and standard deviation from the headlines of the *Wall Street Journal*. All regressions include lagged regression, forecast dispersion, firm size, book-to-market, trading volume, $CAR_{-252,31}$, $CAR_{-30,3}$, and $CAR_{-2,-2}$ control variables. The robust t-statistics are in parentheses.

| Parameter | Tetlock | Vader | $Vader_{Directional}$ | Space | SpaceVader |
|---|---|---|---|---|---|
| Sentiment Mentric | 0.0018*** | 0.0002 | -0.0000 | 0.0004 | 0.0003** |
| | (2.67) | (0.18) | (-0.09) | (1.65) | (2.11) |
| $SUE_{lag}$ | -0.001 | 0.0000 | -0.0001 | -0.0001 | -0.0001 |
| | (-0.90) | (-0.89) | (-0.09) | (-0.89 ) | (-0.89) |
| Forcast Dispersion | 0.0018 | 0.0038 | 0.0018 | 0.0017 | 0.0018 |
| | (0.47) | (0.48) | (0.48) | (0.48) | (0.48) |
| Firm Size | -0.003*** | -0.0009*** | -0.0029*** | 0.0029*** | -0.0029*** |
| | (-3.19) | (-3.08) | (-3.09) | (-3.08) | (-3.08) |
| Book-to-Market | 0.1654 | 0.1522 | 0.1524 | 0.156 | 0.1579 |
| | (0.654) | (0.42) | (0.42) | 0.43) | (-43) |
| TradingVolume | -0.012 | -0.0116 | -0.0116 | -0.0116 | -0.0117 |
| | (-1.47) | (-1.41) | (-1.41) | (0.1635) | (-1.41) |
| $CAR_{-252,-31}$ | -0.0003 | -0.0002 | -0.0003 | -0.0003 | -0.0003 |
| | (-0.13) | (-0.13) | (-0.13 | (-0.14) | (-0.14) |
| $CAR_{-30,-3}$ | 0.0002 | 0.0000 | -0.0002 | -0.0001 | -0.0001 |
| | (0.02) | (0.00) | (0.01) | (-0.01) | (-0.01) |
| $CAR_{-2,-2}$ | -0.004 | -0.0029 | -0.0029 | -0.0027 | 0.0027 |
| | (-0.10) | (-0.07) | (-0.07) | (-0.07) | (-0.07) |
| Observations | 14008 | 14008 | 14008 | 14008 | 14008 |
| Clusters | 76 | 76 | 76 | 76 | 76 |
| $Adjusted R^2$ | 0.0018 | 0.0015 | 0.0010 | 0.0012 | 0.0011 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

throughout everyday events. Using daily returns following WSJ articles, I measure the impact of all five sentiment metrics on daily headlines. As the time of each article is not available through the WSJ archive, we measure the article sentiment one day before the following abnormal return.

In each regression I include several control variables following Tetlock closely. I include the firms most recent earnings surprise (SUE), along with abnormal returns for $AR_{-1,-1}$ one day before the story and $AR_{-2,-2}$ and two days before the story as well as one abnormal return for the previous calendar year $CAR_{-252,-31}$. I also control for firm size, book-to-market, and trading volume using the same methods mentioned in the previous section. I use a pooled regression and standard errors clustered by time period using the same methodology and for the same reasons as explained above. Table 7 reports the results of the effects of daily headline sentiment on abnormal return the following day.

We can see from Table 7 that most of the sentiment analysis variables are significant. The only significant control is the prior quarters earnings surprise. Tetlock's measure of sentiment is highly significant and negative implying that the more negative a story is, the more negative the subsequent return will be. This is inline with Tetlock's paper and consistent with the logic of negative news preceding negative returns.

We can also see that $Vader_{Directional}$, Space, and SpaceVader are all highly significant. They also have a higher $R^2$ than the traditional negative word count from Tetlock. Furthermore they are able to calculate abnormal price response for both positive and negative news. We can see that the strongest predictor of abnormal return is SpaceVader which is also the most complicated of the other metrics. This does not mean that the more complicated a metric is, the more successful. Instead, it leads to the idea that the information in a headline is significant to investors and that this information can be extracted a myriad of ways.

An additional robustness check was done to compare companies that were frequently mentioned in the news and those that were not. I created one variable which measured the abnormal mentions in the news which was the number of mentions in the [-30, -3] window standardized by the prior years average and standard deviation. This variable did not have strong significance and was not included. I also seperated the companies into quartiles based on average annual mentions in the news and repeated the above mentioned regressions. There was no significant change in results and so this was not included.

Table 7: This table compares five separate sentiment metrics against the Abnormal Return (AR) 1 day after the *Wall Street Journal* article headline is published. The independent variable is the abnormal return and is calculated as the excess return of the S&P 500. The regression is an ordinary least squares regression clustered around calendar quarters. The sentiment metrics are calculated as the sentiment score of all the company specific headlines on one day standardized by the previous years average and standard deviation. All regressions include previous earnings surprise (SUE), firm size, book-to-market, trading volume, $CAR_{-252,31}$, $AR_{-2,-2}$, and $AR_{-1,-1}$ control variables. The robust t-statistics are in parentheses.

| Parameter | Tetlock | Vader | $Vader_{Directional}$ | Space | SpaceVader |
|---|---|---|---|---|---|
| Sentiment Mentric | -0.0005*** | -0.0000 | -0.0000*** | 0.0004*** | 0.0008*** |
| | (-2.72) | (-0.65) | (-25.73) | (3.44) | (6.12) |
| SUE | -0.0001** | -0.0001** | -0.0001** | -0.0001** | -0.0001** |
| | (-2.03) | (-1.97) | (-1.97) | (-1.99 ) | (-2.04) |
| Firm Size | -0.0001 | -0.0009 | -0.0001 | -0.0001 | -0.0001 |
| | (-3.19) | (-0.22) | (-0.22) | (-0.20) | (-0.18) |
| Book-to-Market | -0.2369 | -0.2353 | -0.2357 | -0.2382 | -0.2379 |
| | (-1.27) | (1.26) | (-1.26) | (-1.27) | (-1.27) |
| TradingVolume | 0.0001 | 0.0002 | 0.0002 | 0.0004 | -0.0004 |
| | (0.01) | (0.02) | (0.02) | (0.06) | (0.07) |
| $CAR_{-252,-31}$ | 0.0002 | 0.0003 | 0.0003 | 0.0002 | 0.0001 |
| | (0.25) | (0.28) | (0.28) | (0.22) | (0.13) |
| $AR_{-1,-1}$ | -0.011 | -0.0106 | -0.0106 | -0.111 | -0.0172 |
| | (-.81) | (-0.80) | (-0.80) | (-0.84) | (-1.14) |
| $AR_{-2,-2}$ | -0.0168 | -0.0165 | -0.0164 | -0.017 | 0.0.0179 |
| | (-1.11) | (-.80) | (-1.09) | (-1.11) | (-0.92) |
| Observations | 34937 | 34397 | 34397 | 34397 | 34397 |
| Clusters | 5746 | 5746 | 5746 | 5746 | 5746 |
| $AdjustedR^2$ | 0.0008 | 0.0055 | 0.0006 | 0.0011 | 0.002 |

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

It is clear that these metrics, when used to estimate next day returns, do not offer a profitable trading strategy as the coefficients are near or equal to zero. Instead, they offer insight into the potential of computational linguistics as a tool to understand irrational behavior. From these results, we see that the language of a headline has indeed an effect on how investors will perceive their investments. But this is only a first step. There is much more work to do, and much more potential to develop language analysis techniques that can capture behavioral traits not readily visible to a human reader. Potentially, long term price patterns not measured here could be affected by news. Or perhaps a broader range of news sources is necessary to return more significant results.

# 7 Conclusion

Exactly how to quantify sentiment in news has been shown here to be a precarious process. Methodologies for measuring sentiment should be changed depending on source and intent. However, what I have shown is that there is more to news than word counts, especially just negative counts. The complexity of language cannot be ignored and should be included in an in depth financial analysis. I have confirmed, as Tetlock did, that there is a subjective, qualitative reaction to news and this finding is not inline with rational investors, or market efficiency. This study of language scratches the surface of the complicated relationship between perception and asset pricing. The language in news sends a signal that investors interpret in complicated ways beyond what linear counts of negative words can capture. Including lexical analyzers, customized dictionaries, and parts of speech analyzers does not resolve the mystery of this relationship but further illustrates its complexity. A more in depth analysis, with more advanced knowledge of linguistics would potentially disambiguate part of this relationship. By examining a larger sample of headlines more closely it may be possible to apply a larger set of language rules that work together to clarify the relationship between perception and languge.

# 8 Works Cited


Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*, 1645-1680.

Barber, B., Odean, T. (2008), All that Glitters: The effect of attention and news on the buying behavior of individual and institutional Investors. *Review of Financial Studies, 21, issue 2*, 785-818.

Basu, S. (1997) The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics, 24*, 3-37.

Bernard, V., Thomas, J. K. (1990). Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics, Volume 13, Issue 4*, 305-340.

Chan, W. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics. 70.*

Da, Z., Engelberg, J. and Gao, P. (2011). In search of attention. *The Journal of Finance, 66*, 1461-1499.

Dellavigna, S. and Pollet, J.M. (2009). Investor inattention and friday earnings announcements. *The Journal of Finance, 64*, 709-749.

DeVault, L., Sias, R. and Starks, L. (2019). Sentiment metrics and investor demand. *The Journal of Finance, 74*, 985-1024.

Engelberg, J., McLean, R.D. and Pontiff, J. (2018). Anomalies and news. *The Journal of Finance, 73*, 1971-2001.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25(2)*, 383-417.

Fama, E, (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics, 49, issue 3*, 283-306.

Fang, L. and Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance, 64*, 2023-2052.

Frazzini, A. (2006). The disposition effect and underreaction to news. *The*

*Journal of Finance, 61*, 2017-2046.

Frieder, L., & Subrahmanyam, A. (2005). Brand perceptions and the market for common stock. *The Journal of Financial and Quantitative Analysis, 40(1)*, 57-85.

Garcia, D. (2013), Sentiment during recessions. *The Journal of Finance, 68*, 1267-1300.

Hartzmark, S.M. and Shue, K. (2018). A tough act to follow: Contrast effects in financial markets. *The Journal of Finance, 73*, 1567-1613.

Hirshleifer, D., LIM, S.S. and TEOH, S.H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance, 64*, 2289-2325.

Hirshleifer, D., Myers, J., Myers, L., & Teoh, S. (2008). Do individual investors cause post-earnings announcement drift? Direct evidence from personal trades. *The Accounting Review, 83(6)*, 1521-1550.

Jiang, F., Lee, J., & Xiumin, M. & Guofu, Z. (2019). Manager sentiment and stock returns. *Journal of Financial Economics, Elsevier, vol. 132(1)*, 126-149.

Kahneman, D., Tversky, A. (1990). *Prospect theory: An analysis of decision under risk. Cambridge University Press*, 263-91.

Kahneman, D., Tversky, A. (1990). Prospect theory: An analysis of decision under risk. *Cambridge University Press*, 263-91.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and biases*, 1124-1131.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127(2)*, 267–286.

Loughran, T., McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*, 35-65.

Merton, R.C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance, 42*, 483-510.

Mian, G., Sankaraguruswamy, S. (2012). Investor sentiment and stock narket response to earnings news. *The Accounting Review, 87(4)*, 1357-1384.

Neumann, John von and Morgenstern, Oskar, *Theory of games and economic behavior*. *Princeton, NJ. Princeton University Press*, 1953.

Shefrin, H., Statman, M. (1985). The disposition to sell winners too early and ride Losers too long: Theory and evidence. *The Journal of Finance, 40(3)*, 777-790.

Simon, H. A. (1957). *Models of man: Social and rational; mathematical essays on rational human behavior in society setting*. *New York: Wiley*.

Schumaker, R. and Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, Art. 12.

Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*, 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008). More than words: quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*, 1437-1467.

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science. 4 (3)*, 199–214.

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review, 106 (7)*, 1577-1600.

Tversky, A., Kahneman, A. (1973). Availability: A heuristic for judging frequency and probability, *Cognitive Psychology, Volume 5, Issue 2*, 207-232.

Tversky, A., Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185(4157)*, 1124-1131.

Varian, Hal. (1985). Non-parametric analysis of optimizing behavior with measurement error. *Journal of Econometrics, 30, issue 1-2*, 445-458.

Weisbrod, E. (2019). Stockholders' unrealized returns and the market reaction to dinancial disclosures. *The Journal of Finance, 74*, 899-942.

Yuan, Y. (2015). Market-wide attention, trading, and stock returns. *Journal of Financial Economics, 116, issue 3*, 548-564.