Summer 5-15-2023

Edgar 10X filings’ sentiment and predicting a firm’s stock return

James Horn
Washington University in St. Louis, jjphorn@gmail.com

WASHINGTON UNIVERSITY IN ST. LOUIS

Olin School of Business

Dissertation Examination Committee:
Guofu Zhou, Chair
Xing Huang
Xiumin Martin

Edgar 10X Filings' Sentiment and Predicting a Firm's Stock Return

by
James F. Horn

A dissertation presented to
the Olin School of Business
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctorate of Business Administration

August 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I appreciate Guofu Zhou's guidance and patience, he is a great mentor and I am thankful for the countless hours of him responding to questions. I am forever grateful to Majeed Simaan for teaching me how to combine SEC data with CRSP/COMPUSTAT. I am forever indebted to Josh Ulrich for providing me with a server and expert coding assistance. Finally, I wouldn't have been able to figure out some of the complex coding required if it wasn't for KaiHong Luo, his work ethic is admirable.

James F. Horn

Washington University in St. Louis
August 2023

Dedicated to my children: JP, Christopher, Julianna, George and Genevieve.

ABSTRACT OF THE DISSERTATION

Edgar 10X Filings' Sentiment and Predicting a Firm's Stock Return

by

James F. Horn

Doctorate of Business Administration in Finance

Washington University in St. Louis, 2023

Professor Guofu Zhou, Chair

This research paper investigates the role of managerial sentiment in predicting future stock returns by constructing firm characteristics based on sentiment words. Through the analysis of all Securities and Exchange filings submitted by publicly traded companies from 2001 to 2022, this study demonstrates that certain sentiment words provide a robust and significant cross-sectional predictor of future monthly stock returns. The paper utilizes a unique approach by leveraging a long time frame and analyzing all available filings. This allows for a comprehensive analysis of the impact of managerial sentiment on stock returns, highlighting the importance of sentiment analysis in financial markets.

Chapter 1

Introduction

A large literature in behavioral finance examines the effect of sentiment on asset prices. Much of this literature studies the effect of investor sentiment on asset prices. Excess investor optimism and pessimism can drive a wedge between an asset's price and present value. [3] develop a theoretical model showing investor sentiment may push an asset's price away from the asset's fundamental. [1] constructs an empirical index measuring investor sentiment. [6] also constructs an index from several proxy variables measuring investor sentiment. A recent review on investor sentiment is [12]. [5] shows model-based estimates of expected returns and investors' expectations of future stock returns are negatively correlated. There is less research on the relationship between manager sentiment and asset prices.

In this dissertation, I consider manager sentiment a plausible source of variation in asset prices because managers determine a firm's operations, investments, and payouts to investors. Managers also provide information to investors regarding firm operations. Manager sentiment may differ from investor sentiment because managers hold private information about their firms operations. [9] shows managerial overconfidence distorts corporate investment. [2] reviews research on managers' behavioral biases. [7] show an index of aggregate manager sentiment negatively forecasts future aggregate stock returns.

This research paper investigates the role of managerial sentiment in predicting future stock returns by constructing firm characteristics based on sentiment words. Through the analysis of all Securities and Exchange filings submitted by publicly traded companies from 2001 to 2022, this study demonstrates that certain sentiment words provide a robust and significant cross-sectional predictor of future monthly stock returns. The paper utilizes a unique approach by leveraging a long time frame and analyzing all available filings. This allows for a comprehensive analysis of the impact of managerial sentiment on stock returns, highlighting the importance of sentiment analysis in financial markets. The findings of this study have significant implications for investors, as they suggest that incorporating sentiment analysis into investment strategies can help identify potentially lucrative investment opportunities. Additionally, this study provides important insights into the role of managerial sentiment in financial markets, shedding light on the underlying factors that drive stock returns. Overall, this paper contributes to the growing body of literature on sentiment analysis and its impact on financial markets. The use of a comprehensive dataset and a long time frame enhances the validity and generalizability of the findings, providing a valuable resource for academics and practitioners alike.

In contrast to existing studies, I specifically examine the cross-sectional relationship between managers' sentiment and US stock returns. I focus on managerial sentiment for a few reasons. First is that information, fundamental or sentimental, annual filings provide about these facets of a business's operations will influence the firm's profitability, and hence the firm's share price accordingly. Secondly, Managerial sentiment may also influence firm value by influence a manager's choices regarding corporate financing, investment, and shareholder payouts. Third, available empirical studies indicate manager sentiment is an economically important source of variation in firm-level and aggregate stock returns. [8] shows manager sentiment predicts stocks' excess returns for five days following a filing's release to the public. [12] shows aggregate manager sentiment is a strong time series predictor of future aggregate stock market returns.

I construct a measure of manager sentiment from all of the 10Xfilings. I use the 10X filings to construct a text-based measure of sentiment because the filings' description of firms' past and future operations contain some information about managers' opinions, judgements, and beliefs regarding firms' past and future economical fundamentals I measure the sentiment of a firm's manager by counting the occurrence of each positive and negative word from [8]. My approach is very similar to the standard dictionary approach for constructing text-based variables from word lists. The difference is I focus on the predictability of each word and not the entire list.

Studies using the standard, related dictionary method to measure document done and sentiment are [11], [8], [4], and [7]. My work differs from these studies in a few important ways. First, I independently evaluate all the positive and negative words to see their effects on stock returns. Second, I include a broader range of a historical time span, to include 2003, a commonly. dropped year, to test for variable robustness. Third, I include all 10X filings.

The findings of this study have significant implications for investors, as they suggest that incorporating sentiment analysis into investment strategies can help identify potentially lucrative investment opportunities. Additionally, this study provides important insights into the role of managerial sentiment in financial markets, shedding light on the underlying factors that drive stock returns.

The rest of the paper is organized as follows. In Section 1.2, I review the construction of the word count variables. In Section 1.2, I discuss the regression methodology used to identify the predictability of each word count variable. In Section 1.3, I present the results of the model. Section 1.4 concludes.

1.1 Data

1.1.1 Data Sources

To build the complete data set for my analysis, I combine data from the following three sources: Compustat, the Center for Research in Security Prices (CRSP), and all the text filings for the 10-Ks, 10-Qs and their variants from Bill McDonald's website. I follow most of steps outlined in [10] to filter and merge the datasets. Starting with CRSP, I download the permanent identifier of the security (PERMCO), date, shares outstanding in thousands (SHROUT), share code (SHRCD), exchange code (EXCHCD), security return (RET), price (PRC), and the CUSIP identifier. The key thing to point out is that the CUSIP is the key link between the CRSP and COMPUSTAT data.

10X Filings

Definition

Throughout this paper, I refer to all filings as "10-X". This represents any Securities and Exchange (SEC) filing that is a 10-K variant, e.g., 10-Q, 10-K/A, 10-K405, etc. Please see Table 1 for a complete breakdown of the filing types and counts for the sample period. These annual and quarterly filings are required by any issuer with securities registered under Section 12 or subject to Section 15(d) of the SEC of the Securities Exchange Act of 1934, as amended, and subject to the periodic and current reporting requirements of Section13 or 15(d). For a in depth description of the timing and dissemination of SEC filings, please see (need to cite) is provided by Rogers, Skinner, and Zechman (2017, Journal of Accounting Research).

Table 1.1: All Edgar Filings - 2001:2022
This table shows the breakdown of the different filing types.

Filing Type	Sample Size
EDGAR 10-K/10Q complete sample	710,479
10-K	168,444
10-K-A	33,727
10-K405	5,168
10-K405-A	617
10-KSB	1
10-KT	468
10-KT-A	114
10-Q	469,627
10-Q-A	32,143
10-QSB	4
10-QSB-A	1
10-QSB-A	1
10-QT	147
10-QT-A	16

1.1.2 Data Screens

I filter the CRSP data by doing some of the accepted best practices in financial research. I start by keeping on the common shares, those stocks with SHRCD 10 or 11 codes. I then drop any observations that are missing values for prices. Furthermore, I keep securities listed on major exchanges (NYSE, AMEX, or NASDAQ). After applying these filters, I then compute the market cap for each stock-month in the data. Since I link the data with COMPUSTAT using the CUSIP, I check whether there are duplicates for the same cusip-month and drop any duplicate values. I also require a minimum history of at least 10 months of data, since the smooth beta I use for forecasting requires 4 months of data. To avoid survivorship-bias, especially when it comes to researching the cross-section of stock returns, I drop all companies that were delisted in the past.

To control for delisted returns, I follow the methodology recommended by Bali et al 2016.

For Compustat, I begin with all observations in the Compustat North America Fundamentals Quarterly database between January 1, 2001 and December 31, 2022. I omitted firms that did not have a CIK or CUSIP. The CIK is key since it links firms via the SEC Edgar filings. I then merged CRSP and Compustat which resulted in having 8,423 unique Cusips.

1.1.3 Preprocessing

The 10-K documents were parsed using the methods described in Loughran and McDonald (2014). I did not do common textual pre-processing steps such as stemming, removing numbers, removing punctuation and removing stop words. None of these would have an impact on our ability to search for and subset the sentences containing the sentiment words.

1.1.4 Parsing

For the parsing step, I individually imported each filing since there were 710,479 filings in scope. Once I imported a filing, I split each filing into separate sentences by using know sentence splitting techniques, such as splitting a sentence by a period followed by a space. I then stored each sentence into its own vector per filing and then combined them into a data frame. This gave us one data frame per filing which allowed us to start looking for the sentences that contained the word growth. I then checked for the presence of the word growth and only keep the sentences that contain the word growth.

In this section, I discuss the key areas associated with the implementation of textual analysis and consider various aspects of executing our method. Our aim is to make our research easy to reproduce, which can be a tough task in textual analytics.

Defining a Word

Since our analysis focuses on a word, in this case the word growth, I find it necessary to define the term word. Since almost all text analytic methods are based on first identifying words, one of the first steps is to parse each document into a vector of tokens, where tokens are collections of characters occurring between word boundaries. In our case, a word boundary is a white space. I removed all extra white space in the document before parsing so that only single whites paces remained. This step can produce a relatively long list which is substantially and meaningfully shortened by selecting only those tokens which map into a list of words. This, in turn, requires the researcher to specify what collections of characters are considered words.

In our case, our approach is straightforward in that I have a target phrase which is well defined, is not likely to be misidentified (e.g., it usually does not show up in filing headers or as a homonym), and isn't affected by morphemes. This makes it a very easy word to parse out of a document.

Focusing on a target word is one of the simplest and most effective approaches to textual analysis. It avoids the challenge of ambiguity, large word lists are much more prone to error when compared to tests focusing on a few unambiguous words or phrases. An example of focusing on a word or target phrase is Loughran, McDonald, and Yun [2009] research that looks at the frequency of the word "ethic" (and its variants) along with the phrases "corporate responsibility", "social responsibility", and "socially responsible" in 10-K filings. These measures were used to determine if these words were associated with with "sin" stocks, corporate governance measures, and class action lawsuits. In their results, they find that firms whose managers focused on using these phrases were more likely to be labeled as sin stocks, have low corporate governance measures, and be sued in the year subsequent to the filing.

The parsing required to achieve results such as these is relatively straightforward and easily replicated. In a subsequent section, I provide an example of textual analysis using this simple context where I look for the specific sentiment word in financial disclosures.

Defining a Sentence

A question that I want to answer to help provide clarity in our research is: What is a sentence? (Loughran Mcdonald 2016) Since sentences are used in well-known research instances, specifically the Fog Index, I want to clarify what I mean by a sentence, how I parse it, and how I calculate the words per sentence. Generally the first step before analyzing sentences, researches will remove abbreviations, headings, and numbers (with decimals), and then assume the remaining periods are sentence terminations. Average words per sentence is then determined by the number of words divided by the number of sentence terminations. While this does provide an average number of words per sentence, it can lead to significant parsing errors. Bushee, Gow, and Taylor [2015] highlight that the Perl routine used by Li [2008] to calculate the number of words per sentence is "confounded by punctuation used in numbers and abbreviations." This can lead researchers to misidentify sentence terminations and cause parsing errors that are extraordinary.

To avoid the volatility of using the average sentence length, I only parsed the sentences that contained the word growth. This allowed us to avoid the volatility of using the average sentence length as a key measure. While are approach is not completely error prone, it reduces the risks of errors and is easily repeatable.

Defining our Corpus

One corpus of textual material that has received considerable attention is MD&A section of the 10-K filing. While this section is generally the core section of most analysis, I am using the entire filing. In summary, I look at the business section, the F-pages, the Risk Factors, and the MD&A. For example, the business section contains a significant amount of wording that comes from individuals outside of the accounting and finance department. The word growth appears in many aspects of the filings where managers may use the term so I did not confine our search to a particular segment. I am looking at the overall use of the word.

Variable creation

The word count variables were constructed by taking the total sentiment word count per filing and dividing it by the total words in the filing. This allows for the normalizing of the ratios due to filing types being influenced by the size of the firm.

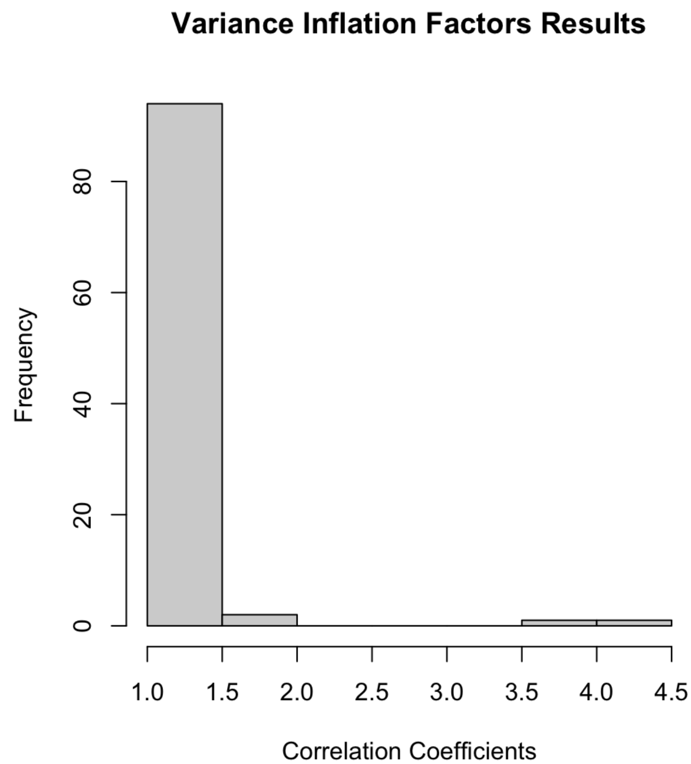

Figure 1.1: Variance inflation factor.

Correlations amongst word groups

Since I have 97 word count variables, it is important to address the concern of multicollinearity. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with other. My goal for testing for multicollinearity are twofold: One, isolate the independent variables and the effects they have on the dependent variables, and two, to make sure that when I combine the top variables together that I'm not combining correlated variables.

One common way to detect multicollinearity is by using the variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model. It is calculated by taking the the ratio of the variance of all the independent variables betas divide by the variance of a single beta if it were fit alone.

The median VIF is 1.16, with two that right skew it. Generally the cut off for excluding variables is above 7, so the two that range from 3.5 to 4.5 are not a significant concern for multicollinearity.

1.2 Regression Model

Using cross-sectional regression (CSR) to identify factor models is well documented and well known in finance literature. CSR models the relationship between a dependent variable and one ore more independent variables. The independent variables are represented as factors.

The general form of a cross-sectional regression factor model is:

$$Y = \beta_0 + \beta_1 XF1 + \beta_2 XF2 + \cdots + \beta_k XFK + \epsilon, \tag{1.1}$$

where:
Y: the dependent variable
XFi: the factors
β_i: the coefficients representing the impact of each factor
ϵ: error term

To identify the top variables that have predictive power, I used the rolling cross-section regression (RCSR). It is a variation of a standard cross-sectional regression, where the regression is performed on a panel dataset with a fixed set of independent variables, but with rolling periods of time, where each regression is conducted using a subset of the most recent observations. The window of observations used in the regression is then rolled forward, and the regression analysis is repeated on the new subset of the most recent observations. This process continues until all observations have been included in at least one regression.

I chose this model as it is a useful technique to model the dynamic relationships amongst variables over time, and can provide insights into patterns and trends that my not show in a traditional cross-sectional regression.

RCSR can be used to test for changes in the relationship between two variables over time, or to identify periods when the relationship is stronger or weaker. It can also help to avoid issues with time-varying coefficients and heteroscedasticity that may be present in standard cross-sectional regressions.

$$\begin{pmatrix} R_{1,t} \\ R_{2,t} \\ . \\ . \\ . \\ R_{1000,t} \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,4} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,4} \\ . & . & . & . & .\cdots \\ . & . & . & . & .\cdots \\ . & . & . & . & .\cdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,98} \end{pmatrix} \begin{pmatrix} \beta_{1,t} \\ \beta_{2,t} \\ . \\ . \\ . \\ \beta_{n,t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ . \\ . \\ . \\ \varepsilon_{n,t} \end{pmatrix}$$

Where:

R is an $n \times 1$ matrix of the dependent variable
X is an $n \times 97$ matrix of the independent variables, with the first column consisting entirely of ones to account for the intercept term
β is a 98×1 matrix of the regression coefficients, with the first element corresponding to the intercept term and the remaining 97 elements corresponding to the coefficients of the independent variables
ϵ is an $nx1$ matrix of the error terms.

The general formula for a multi-period cross-sectional regression is similar to that of a single-period cross-sectional regression, but it includes an additional subscript for time:

$$Y_{it} = \beta_0 + \beta_1 X1_{it} + \beta_2 X2_{it} + \cdots + \beta_k Xk_{it}\epsilon_{it} \tag{1.2}$$

The first portfolios I create are equal weighted using OLS and creating portfolios based off the regression coefficients. Table 2 shows the results for the top 10 word equal weighted portfolios. In Table 3 I show the results for combining the top 3, 5, 10 and all variables and rerun the RCRS/OLS regression to see if there is an improvement in performance. Table 4 shows the results for combining the top word counts. To Do: Put formulas in text. Also fix Table 5. look at table numbers.

Chapter 2

Portfolio Results

In this chapter I show the results of creating the word count variables by using the Equal-weighted and value-weighted methods. Each method is a completely different way of constructing a portfolio, and each has its advantages and disadvantages.

Equal-weighted portfolios allocate an equal amount of money to each security in the portfolio. This type of portfolio construction gives each security an equal weighting, regardless of its market capitalization or valuation. For example, if an investor had $1000 to invest in 10 stocks they would buy $100 worth of each stock. This approach can be useful in avoiding the concentration of risk in a small number of high-value stocks, and it can potentially lead to higher returns if the smaller stocks in the portfolio outperform the larger ones.

On the other hand, value-weighted portfolios allocate a larger percentage of the portfolio to securities with higher market capitalizations or valuations. This approach gives more weight to stocks that are more valuable in the market. This approach can be useful in capturing the performance of the overall market, and it can potentially lead to more stable returns over time.

I create both equal weighted and value weighted portfolios below to gain insight into the behaviors of the different weighted portfolios. I create four tables for each weighted portfolio method, I first show the performance results of all variables for the first and second half, then I show the performance of the Top 10 variables in each half, which are sorted by the absolute value of the Sharpe ratio. After that I show the results for investing in the Top 10 variables in the first half to see their out of sample return, which is the second half. Finally I combine the Top 3, Top 5, and Top 10 variables by absolute Sharpe ratio from the first half and and show their second half performance.

2.0.1 Equally Weighted Portfolios

An equal weighted portfolio is a type of portfolio where each stock in the portfolio is allocated an equal weight of the overall portfolio value. What this means is that every stock in the portfolio has the same impact on the portfolio's overall performance. This type of weighted portfolio is used as a way to diversify risk since equal weight is given to each stock in the portfolio. This means that the portfolio's performance is not overly concentrated on one stock and typically more stocks are held. If one stock performs better than the others and its value increases, it will have the same impact on the portfolio's overall performance as any other stock. Additionally, equal weighted portfolios can provide exposure to a wide range of industries and sectors, which can further diversify the portfolio. It's worth noting that equal weighted portfolios will not necesarly outperform a value weighted portfolio, in all market conditions.

The first portfolios I create are equal weighted portfolios using RCSR with OLS. I then sort the portfolios into ten groups based of the regression coefficients. From there I take the Long minus Short portfolios. There are a total of 97 word count portfolios.

Table 2.1: All Word Count Portfolios - Equal Weighted
Word Count Portfolios are alphabetically sorted. Results are split into two perdios: First Half and Second Half

	First Half				Second Half			
	avg	std	t-stat	Shp	avg	std	t-stat	Shp
able	-0.61	5.4	-0.37	-0.11	0.58	4.7	0.4	0.12
abundant	-1.25	5.34	-0.76	-0.23	1.94	4.58	1.37	0.42
accomplish	2.87	4.69	1.98	0.61	1.25	4.71	0.86	0.27
achieve	-8.02	12.74	-2.04	-0.63	-8.62	15.75	-1.77	-0.55
advan	-0.18	6.6	-0.09	-0.03	0.96	7.57	0.41	0.13
alliance	-2.64	9.45	-0.91	-0.28	-7.75	10.63	-2.36	-0.73
assure	-5.92	10.46	-1.83	-0.57	-3.83	7.65	-1.62	-0.5
attain	-0.72	4.74	-0.49	-0.15	0.84	4.24	0.64	0.2
attractive	-1.3	5.58	-0.76	-0.23	-0.97	5.22	-0.6	-0.19
benefit	-6.51	5.86	-3.6	-1.11	-5.31	9.27	-1.86	-0.57
best	-0.36	5.36	-0.22	-0.07	1.4	4.63	0.98	0.3
better	-0.92	5.36	-0.56	-0.17	1.44	4.63	1.01	0.31
bolster	-0.79	5.06	-0.51	-0.16	0.6	4.35	0.45	0.14
boom	-0.96	5.66	-0.55	-0.17	0.12	4.73	0.08	0.02
boost	-1.21	5.2	-0.75	-0.23	-0.18	4.56	-0.13	-0.04
breakthrough	-1.58	5.55	-0.92	-0.28	0.38	4.81	0.25	0.08
buy	0.99	4.96	0.64	0.2	0.01	6.29	0	0
collaborate	-6.07	13.02	-1.51	-0.47	2.4	19.64	0.4	0.12
compliment	-1.94	5.47	-1.15	-0.35	-0.04	4.76	-0.03	-0.01
conclusive	1.96	4.76	1.34	0.41	0.45	5.65	0.26	0.08
confident	-6.05	8.01	-2.45	-0.75	-4.5	8.99	-1.62	-0.5
constructive	-1.58	5.28	-0.97	-0.3	0.94	4.79	0.64	0.2
cultivation	-1.88	5.41	-1.12	-0.35	-0.56	4.73	-0.39	-0.12
delight	-0.79	5.26	-0.49	-0.15	0.48	4.77	0.33	0.1
dependability	-2.31	4.8	-1.56	-0.48	1.09	4.22	0.83	0.26
desirable	-3.98	4.65	-2.77	-0.86	-6	10.42	-1.87	-0.58
diligent	0.55	5.33	0.34	0.1	-0.12	4.84	-0.08	-0.02
distinctions	0.94	4.64	0.66	0.2	1.72	4.74	1.17	0.36

continues on the next page

continues from the previous page

	First Half				Second Half			
	avg	std	t-stat	Shp	avg	std	t-stat	Shp
dream	-1.62	5.22	-1.01	-0.31	0.36	4.69	0.25	0.08
easier	-2.15	6.61	-1.05	-0.33	-0.98	5.95	-0.53	-0.17
efficient	-1.51	5.69	-0.86	-0.26	-2.67	5.64	-1.54	-0.47
empower	-2.59	5.06	-1.66	-0.51	0.89	5.28	0.55	0.17
enable	-5.88	11.35	-1.68	-0.52	-7.49	8.63	-2.81	-0.87
encourag	-3.35	4.2	-2.58	-0.8	-1.83	4.92	-1.2	-0.37
enhance	-2.3	11.12	-0.67	-0.21	-0.58	7.5	-0.25	-0.08
enjoy	-2.04	5.81	-1.14	-0.35	-0.94	5.06	-0.6	-0.19
enthusiasm	-2.19	5.43	-1.31	-0.4	1.14	4.82	0.76	0.24
evolution	0.29	12.8	0.07	0.02	0.99	9.04	0.35	0.11
excellence	-2.88	5.47	-1.7	-0.53	1.41	4.54	1	0.31
exceptional	-1.6	5.45	-0.95	-0.29	0.44	4.67	0.3	0.09
excited	-0.77	5.01	-0.5	-0.15	0.14	5.23	0.09	0.03
exclusive	-8.32	8.65	-3.12	-0.96	-13.27	15.8	-2.72	-0.84
expansion	-0.03	7.7	-0.01	0	-7.15	11.62	-1.99	-0.62
favorite	-4.02	4.88	-2.67	-0.82	4.64	5.01	3	0.92
gain	-2.59	8.23	-1.02	-0.31	-4.12	10.73	-1.24	-0.38
good	3.36	8.91	1.22	0.38	-3.41	8.32	-1.33	-0.41
great	-1.12	6.31	-0.57	-0.18	0.73	5.69	0.42	0.13
growth	1.1	8	0.45	0.14	-2.6	9.26	-0.91	-0.28
happ	0.49	5.43	0.29	0.09	-4.21	6.25	-2.18	-0.67
high	-6.61	6.4	-3.35	-1.03	-4.93	10.4	-1.54	-0.47
hire	1.27	9.03	0.46	0.14	-8.87	9.32	-3.08	-0.95
honor	2.3	4.96	1.5	0.46	-0.2	4.6	-0.14	-0.04
ideal	-0.64	5.1	-0.41	-0.13	1.29	4.63	0.9	0.28
impress	-1.43	5.3	-0.88	-0.27	-0.55	4.83	-0.37	-0.11
improv	-0.9	6.22	-0.47	-0.14	4.43	6.32	2.27	0.7
increas	-5.74	9.55	-1.95	-0.6	-1.48	10.17	-0.47	-0.15
incredibl	-0.87	5.25	-0.53	-0.16	0.09	4.66	0.06	0.02

continues on the next page

continues from the previous page

	First Half				Second Half			
	avg	std	t-stat	Shp	avg	std	t-stat	Shp
innov	-4.18	8.6	-1.58	-0.49	-7.71	7.56	-3.3	-1.02
innovat	-2.37	8.73	-0.88	-0.27	-8.05	7.52	-3.47	-1.07
invent	0.73	9.95	0.24	0.07	-1.25	8.95	-0.45	-0.14
invest	2.01	8.17	0.8	0.25	-1.35	8.53	-0.51	-0.16
lead	-9.13	10	-2.96	-0.91	-0.18	6.57	-0.09	-0.03
loyal	2.87	6.11	1.52	0.47	-0.72	6.72	-0.35	-0.11
new	5.24	5.02	3.38	1.04	2.92	5.33	1.78	0.55
opportunit	-0.89	5.46	-0.53	-0.16	-3.15	6.32	-1.62	-0.5
outperform	-0.62	5.06	-0.39	-0.12	0.9	4.87	0.6	0.18
perfect	-1.86	6.63	-0.91	-0.28	1.3	5.51	0.77	0.24
please	-0.3	4.77	-0.21	-0.06	-1.15	4.49	-0.83	-0.26
popular	-2.26	6.08	-1.2	-0.37	-2.44	6.15	-1.29	-0.4
positive	-1.96	5.66	-1.12	-0.35	2.16	5.92	1.18	0.37
preeminen	-1	5.39	-0.6	-0.19	0.59	4.66	0.41	0.13
premier	-1.17	5.38	-0.7	-0.22	0.65	4.37	0.48	0.15
prestig	-0.81	5.82	-0.45	-0.14	0.02	4.53	0.02	0.01
proactive	-0.94	4.66	-0.66	-0.2	0.52	4.98	0.34	0.1
proficien	-0.98	5.51	-0.57	-0.18	0.27	5.33	0.16	0.05
profitab	1.33	8.18	0.53	0.16	0.42	9.44	0.14	0.04
progress	-1.99	6.38	-1.01	-0.31	1.58	9.03	0.57	0.17
prosper	-0.57	5.38	-0.34	-0.11	1.11	4.58	0.79	0.24
rebound	-2.25	5.42	-1.34	-0.41	1.5	4.5	1.08	0.33
receptive	-1.16	5.28	-0.71	-0.22	0.56	4.7	0.38	0.12
regain	-0.1	5.04	-0.06	-0.02	-0.08	4.05	-0.06	-0.02
resolve	-2.43	4.58	-1.71	-0.53	1.45	4.35	1.08	0.33
revolutionize	-1.02	5.28	-0.62	-0.19	0.33	4.76	0.22	0.07
reward	-3.08	4.95	-2.02	-0.62	-0.15	5.72	-0.09	-0.03
satis	0.29	5.7	0.16	0.05	-2.86	6.47	-1.43	-0.44
smooth	-0.81	5.31	-0.49	-0.15	0.24	4.39	0.18	0.06

continues on the next page

continues from the previous page

	First Half				Second Half			
	avg	std	t-stat	Shp	avg	std	t-stat	Shp
solv	2.23	5.19	1.4	0.43	-0.17	4.3	-0.13	-0.04
stabili	-2.92	6.54	-1.45	-0.45	0.36	5.15	0.22	0.07
strength	-0.66	4.85	-0.44	-0.14	-0.23	4.71	-0.16	-0.05
succe	0.23	13.01	0.06	0.02	-12.09	16.54	-2.37	-0.73
superior	-1.76	6.14	-0.93	-0.29	-4.25	7.89	-1.74	-0.54
surpass	-0.29	5.06	-0.19	-0.06	0.89	4.55	0.63	0.2
upturn	-1.24	5.6	-0.72	-0.22	1.2	5.2	0.75	0.23
valuable	-1.15	5.23	-0.71	-0.22	0.65	4.73	0.44	0.14
versatil	-0.98	4.92	-0.65	-0.2	1.54	4.45	1.12	0.35
vibra	-1.93	5.33	-1.17	-0.36	0.69	4.42	0.51	0.16
win	-2.27	4.14	-1.78	-0.55	0.37	5.28	0.23	0.07

Next in Table 2.2 I look at the performance of the top 10 variables from the first half and the second half to get an overall look into the top 10 performance of each half. To select the top 10, I sort by the absolute value of the Sharpe ratio. The exclusive word count portfolio appears in the top 10 for both periods, it has a mean return of -8.32 with a Sharpe ratio for the first half and a return of -13.27 with a Sharpe ratio of .84 for the second half. Reversing the portfolio from Long - Short to Short - Long would make this a profitable portfolio.

Table 2.2: Top 10 Word Portfolios - Equal Weighted
Top 10 equal weighted word count portfolios for the first half period and the second half period. Word portfolios are sorted by absolute Sharpe ratio value. All results are annulaized.

First Half Period					Second Half Period				
word	avg	std	t-stat	Shp	word	avg	std	t-stat	Shp
benefit	-6.51	5.86	-3.6	1.11	innovat	-8.05	7.52	-3.47	1.07
new	5.24	5.02	3.38	1.04	innov	-7.71	7.56	-3.3	1.02
high	-6.61	6.4	-3.35	1.03	hire	-8.87	9.32	-3.08	0.95
exclusive	-8.32	8.65	-3.12	0.96	favorite	4.64	5.01	3	0.92
lead	-9.13	10	-2.96	0.91	enable	-7.49	8.63	-2.81	0.87
desirable	-3.98	4.65	-2.77	0.86	exclusive	-13.27	15.8	-2.72	0.84
favorite	-4.02	4.88	-2.67	0.82	alliance	-7.75	10.63	-2.36	0.73
encourag	-3.35	4.2	-2.58	0.8	succe	-12.09	16.54	-2.37	0.73
confident	-6.05	8.01	-2.45	0.75	improv	4.43	6.32	2.27	0.7
achieve	-8.02	12.74	-2.04	0.63	happ	-4.21	6.25	-2.18	0.67

Next in Table 2.3 I look at the performance of the top 10 variables from the first half and then check their out of sample performance in the second half to get an overall look into the top 10 performance of each half. To select the top 10, I sort by the absolute value of the Sharpe ratio. For word count portfolios that were negative in the first half, I take the short minus long when investing in the second half. For example, the exclusive word count portfolio appears in the top 10 for the first period as a S-L portfolio. In the second half the S-L was used and the word count portfolio returned a positive 13.27%. Reversing the portfolio from Long - Short to Short - Long made this a profitable portfolio.

Table 2.3: Top 10 Word Portfolios - Equal Weighted
Shows the equal weighted word count portfolios for the frist half period and the second half period, which is the out of sample. Words detonated with an asterisk were Short minus Long portfolios and were such for the second half. All results are annualized..

	Full Period				One Half Period			
word	avg	std	t-stat	Shp	avg	std	t-stat	Shp
benefit*	6.51	5.86	3.6	1.11	5.31	9.27	1.86	0.57
new	5.24	5.02	3.38	1.04	2.92	5.33	1.78	0.55
high*	6.61	6.4	3.35	1.03	4.93	10.4	1.54	0.47
exclusive*	8.32	8.65	3.12	0.96	13.27	15.8	2.72	0.84
lead*	9.13	10	2.96	0.91	0.18	6.57	0.09	0.03
desirable*	3.98	4.65	2.77	0.86	6	10.42	1.87	0.58
favorite*	4.02	4.88	2.67	0.82	-4.64	5.01	-3	0.92
encourag*	3.35	4.2	2.58	0.8	1.83	4.92	-1.2	0.37
confident*	6.05	8.01	2.45	0.75	4.5	8.99	1.62	0.5
achieve*	8.02	12.74	2.04	0.63	8.62	15.75	1.77	0.55

Next I look at the performance results of creating the Top 3, Top 5, and Top 10 word count variables by combining the top variables based off absolute Sharpe ratio. Table 2.3 shows the results:

Table 2.4: Top 10 Word Portfolios - Equal Weighted - LM
Top 10 equal weighted word count portfolios for the first half period and the second half period. Word portfolios are sorted by absolute Sharpe ratio value. All results are annulaized.

	Mean (%)	Std dev (%)	tstat	SR
Top 3	4.39	5.01	2.84	0.88
Top 5	5.32	5.6	3.08	0.95
Top 10	4.29	5.33	2.61	0.81

2.0.2 Value Weighted Portfolios

Table 2.5: All Variables - Value Weighted - Linear Regression

	First Half				Second Half				
	avg	std	t-stat	Shp	avg	std	t-stat	Shp	
Word	Mean (able	-3.44	8.2	-1.36	-0.42	-7.38	9.81	-2.44	-0.75
abundant	-3.37	8.09	-1.35	-0.42	-5.79	9.42	-1.99	-0.61	
accomplish	3.47	8.01	1.4	0.43	-0.97	8.57	-0.37	-0.11	
achieve	-7.51	11.83	-2.06	-0.63	-3.04	12.21	-0.81	-0.25	
advan	-4.52	9.9	-1.48	-0.46	-4.77	9.56	-1.62	-0.5	
alliance	-3.91	7.84	-1.62	-0.5	-5.69	7.42	-2.49	-0.77	
assure	-3.11	13.52	-0.75	-0.23	-7.79	14.88	-1.7	-0.52	
attain	-3.11	7.92	-1.27	-0.39	-1.92	6.32	-0.98	-0.3	
attractive	-1.48	8.06	-0.6	-0.18	-6.82	8.23	-2.68	-0.83	
benefit	-3.57	9.96	-1.16	-0.36	-6.05	10.9	-1.8	-0.55	
best	-2.1	7.94	-0.86	-0.26	-6.71	9.57	-2.27	-0.7	
better	-2.76	8.01	-1.12	-0.35	-5.59	9.39	-1.93	-0.6	
bolster	-2.17	7.95	-0.89	-0.27	-6.89	9.32	-2.4	-0.74	
boom	-3.57	8.13	-1.42	-0.44	-7.65	9.76	-2.54	-0.78	
boost	-4.44	7.23	-1.99	-0.61	-6.78	9.14	-2.4	-0.74	
breakthrough	-3.57	8	-1.45	-0.45	-6.05	9.58	-2.05	-0.63	
buy	0.62	8.01	0.25	0.08	-5.47	7.55	-2.35	-0.72	
collaborate	-4.13	11.54	-1.16	-0.36	-2.69	11.46	-0.76	-0.23	
compliment	-3.73	8.17	-1.48	-0.46	-3.97	8.27	-1.55	-0.48	
conclusive	-1.9	8.32	-0.74	-0.23	-7.88	9.15	-2.79	-0.86	
confident	-4.89	7.91	-2	-0.62	-5.35	8.95	-1.94	-0.6	
constructive	-5.13	7.77	-2.14	-0.66	-6.43	9.4	-2.22	-0.68	
cultivation	-4.45	9	-1.6	-0.49	-4.5	9.37	-1.56	-0.48	
delight	-2.96	8.22	-1.17	-0.36	-7.42	9.22	-2.61	-0.81	

continues on the next page

continues from the previous page

	First Half				Second Half			
	avg	std	t-stat	Shp	avg	std	t-stat	Shp
dependability	-4.14	7.83	-1.71	-0.53	-6.03	9.15	-2.14	-0.66
desirable	-2.21	8.3	-0.86	-0.27	-1.64	8.69	-0.61	-0.19
diligent	-4.43	8.21	-1.75	-0.54	-1.12	10.13	-0.36	-0.11
distinctions	-0.55	8.33	-0.22	-0.07	0.23	6.81	0.11	0.03
dream	-3.66	7.92	-1.5	-0.46	-7.55	9.34	-2.62	-0.81
easier	-3.46	9.01	-1.24	-0.38	-2.41	8.81	-0.89	-0.27
efficient	-5.53	8.64	-2.07	-0.64	-6.07	8.97	-2.19	-0.68
empower	-4.72	8.03	-1.9	-0.59	-5.51	8.78	-2.04	-0.63
enable	-5.69	13.59	-1.36	-0.42	-4.95	8.64	-1.86	-0.57
encourag	-0.24	6.59	-0.12	-0.04	0.64	7.07	0.29	0.09
enhance	-2.1	10.74	-0.63	-0.2	-8.07	9.35	-2.8	-0.86
enjoy	1.37	8.76	0.51	0.16	-6.91	7.62	-2.94	-0.91
enthusiasm	-3.91	7.72	-1.64	-0.51	-7.6	9.55	-2.58	-0.8
evolution	1.91	12.37	0.5	0.15	-2.91	9.78	-0.96	-0.3
excellence	-3.07	7.27	-1.37	-0.42	1.05	5.87	0.58	0.18
exceptional	-3.9	8.13	-1.56	-0.48	-7.36	9.87	-2.42	-0.75
excited	-3.05	8.47	-1.17	-0.36	-7.81	8.94	-2.83	-0.87
exclusive	-5.62	10.93	-1.67	-0.51	-1.81	11.07	-0.53	-0.16
expansion	-1.68	10.26	-0.53	-0.16	-10.61	13.22	-2.6	-0.8
favorite	1.17	8.57	0.44	0.14	2.04	6.21	1.06	0.33
gain	1.57	10	0.51	0.16	-3.29	8.85	-1.2	-0.37
good	2.11	10.89	0.63	0.19	5.52	10.4	1.72	0.53
great	0.73	10.55	0.23	0.07	0.27	13.03	0.07	0.02
growth	5.62	13.75	1.32	0.41	-0.58	9.29	-0.2	-0.06
happ	-3.43	6.61	-1.68	-0.52	-2.1	8	-0.85	-0.26
high	2.68	8.97	0.97	0.3	-5.11	8.78	-1.89	-0.58
hire	-2.6	10.85	-0.78	-0.24	-6.1	10.34	-1.91	-0.59
honor	-4.55	7.79	-1.89	-0.58	-3.84	6.99	-1.78	-0.55
ideal	-3.79	7.96	-1.54	-0.48	-7.14	9.73	-2.38	-0.73

continues on the next page

continues from the previous page

	First Half				Second Half			
	avg	std	t-stat	Shp	avg	std	t-stat	Shp
impress	-3.07	8.19	-1.21	-0.37	-4.83	8.14	-1.92	-0.59
improv	2.34	9.15	0.83	0.26	-0.23	6.9	-0.11	-0.03
increas	1.42	9.8	0.47	0.15	1.57	8.61	0.59	0.18
incredibl	-3.77	8.08	-1.51	-0.47	-8	9.76	-2.66	-0.82
innov	-1.9	9.8	-0.63	-0.19	-5.17	8.21	-2.04	-0.63
innovat	-0.34	9.84	-0.11	-0.03	-5.28	8.12	-2.11	-0.65
invent	-1.2	13.94	-0.28	-0.09	-3.45	9.82	-1.14	-0.35
invest	6.13	11.16	1.78	0.55	0.26	9.95	0.09	0.03
lead	-1.89	10.31	-0.59	-0.18	2.92	6.28	1.5	0.46
loyal	-0.8	7.08	-0.37	-0.11	-0.44	7.11	-0.2	-0.06
new	-0.23	7.99	-0.09	-0.03	0.77	7.15	0.35	0.11
opportunit	-1	8.27	-0.39	-0.12	-3.91	7.24	-1.75	-0.54
outperform	-3.41	8.01	-1.38	-0.43	-6.59	9.66	-2.21	-0.68
perfect	-1.35	10.08	-0.43	-0.13	-6.49	10.75	-1.96	-0.6
please	-1.75	7.99	-0.71	-0.22	-0.13	7.82	-0.05	-0.02
popular	-2.92	8.3	-1.14	-0.35	-7.96	8.22	-3.14	-0.97
positive	-0.17	8.13	-0.07	-0.02	-5.34	8.05	-2.15	-0.66
preeminen	-4.1	7.35	-1.81	-0.56	-7.23	9.75	-2.41	-0.74
premier	-3.53	8.08	-1.41	-0.44	-6.78	9.74	-2.26	-0.7
prestig	-4.19	8.1	-1.67	-0.52	-8.35	9.67	-2.8	-0.86
proactive	-5.16	8.66	-1.93	-0.6	-5.35	8.09	-2.14	-0.66
proficien	-4.22	8.03	-1.7	-0.53	-5.12	8.65	-1.92	-0.59
profitab	2.27	7.94	0.93	0.29	-1.82	7.1	-0.83	-0.26
progress	-1.91	8.47	-0.73	-0.22	-1.63	9.08	-0.58	-0.18
prosper	-3.99	7.83	-1.65	-0.51	-6.79	9.95	-2.21	-0.68
rebound	-4.45	8.05	-1.79	-0.55	-7.31	9.49	-2.5	-0.77
receptive	-3.81	8.2	-1.51	-0.46	-7.45	9.86	-2.45	-0.76
regain	-3.34	7.72	-1.4	-0.43	-7.2	9.57	-2.44	-0.75
resolve	-2.05	9.5	-0.7	-0.22	2.36	8.35	0.91	0.28

continues on the next page

continues from the previous page

	First Half				Second Half			
	avg	std	t-stat	Shp	avg	std	t-stat	Shp
revolutionize	-3.62	8.25	-1.42	-0.44	-7.43	9.66	-2.49	-0.77
reward	-2.19	8.03	-0.88	-0.27	-0.11	6.8	-0.05	-0.02
satis	-3.85	8.45	-1.48	-0.46	-3.17	7.48	-1.38	-0.42
smooth	-3.04	7.76	-1.27	-0.39	-7.36	9.65	-2.47	-0.76
solv	0.11	8.28	0.04	0.01	-0.83	8.63	-0.31	-0.1
stabili	-4.05	10.81	-1.21	-0.37	-8.55	8.7	-3.18	-0.98
strength	-2.9	8.19	-1.15	-0.35	3.02	7.67	1.28	0.39
succe	-2.34	10.44	-0.73	-0.22	-3.47	11.09	-1.01	-0.31
superior	-2.48	8.56	-0.94	-0.29	-7.55	9.05	-2.71	-0.83
surpass	-1.79	8.45	-0.69	-0.21	-6.61	9.3	-2.3	-0.71
upturn	-3.91	8.35	-1.52	-0.47	-6.43	9.49	-2.19	-0.68
valuable	-3.34	8.05	-1.34	-0.41	-7.63	9.74	-2.54	-0.78
versatil	-3.59	8.31	-1.4	-0.43	-5.28	9.12	-1.88	-0.58
vibra	-4.05	8.26	-1.59	-0.49	-5.63	9.44	-1.93	-0.6
win	-2.26	10.3	-0.71	-0.22	-1.55	7.27	-0.69	-0.21

Next I look at the performance of the top 10 variables from the first half and the second half to get an overall look into the top 10 performance of each half. To select the top 10, I sort by the absolute value of the Sharpe ratio. The exclusive word count portfolio appears in the top 10 for both periods, it has a mean return of -8.32 with a Sharpe ratio for the first half and a return of -13.27 with a Sharpe ratio of .84 for the second half. Reversing the portfolio from Long - Short to Short - Long would make this a profitable portfolio.

Table 2.6: Top 10 Word Portfolios - Value Weighted - LM
Top 10 equal weighted word count portfolios for the first half period and the second half period. Word portfolios are sorted by absolute Sharpe ratio value. All results are annulaized.

Full Period					One Half Period				
word	avg	std	t-stat	Shp	word	avg	std	t-stat	Shp
constructive	-5.13	7.77	-2.14	0.66	stabili	-8.55	8.7	-3.18	0.98
efficient	-5.53	8.64	-2.07	0.64	popular	-7.96	8.22	-3.14	0.97
achieve	-7.51	11.83	-2.06	0.63	enjoy	-6.91	7.62	-2.94	0.91
confident	-4.89	7.91	-2	0.62	excited	-7.81	8.94	-2.83	0.87
boost	-4.44	7.23	-1.99	0.61	conclusive	-7.88	9.15	-2.79	0.86
proactive	-5.16	8.66	-1.93	0.6	enhance	-8.07	9.35	-2.8	0.86
empower	-4.72	8.03	-1.9	0.59	prestig	-8.35	9.67	-2.8	0.86
honor	-4.55	7.79	-1.89	0.58	attractive	-6.82	8.23	-2.68	0.83
preeminen	-4.1	7.35	-1.81	0.56	superior	-7.55	9.05	-2.71	0.83
invest	6.13	11.16	1.78	0.55	incredibl	-8	9.76	-2.66	0.82

Table 2.7: Top 10 Word Portfolios - Value Weighted - LM
Shows the value weighted word count portfolios for the frist half period and the second half period, which is the out of sample. Words detonated with an asterisk were Short minus Long portfolios and were such for the second half. All results are annualized..

	Full Period				One Half Period			
word	avg	std	t-stat	Shp	avg	std	t-stat	Shp
constructive*	-5.13	7.77	-2.14	0.66	6.43	9.4	2.22	0.68
efficient*	-5.53	8.64	-2.07	0.64	6.07	8.97	2.19	0.68
achieve*	-7.51	11.83	-2.06	0.63	3.04	12.21	0.81	0.25
confident*	-4.89	7.91	-2	0.62	5.35	8.95	1.94	0.6
boost*	-4.44	7.23	-1.99	0.61	6.78	9.14	2.4	0.74
proactive*	-5.16	8.66	-1.93	0.6	5.35	8.09	2.14	0.66
empower*	-4.72	8.03	-1.9	0.59	5.51	8.78	2.04	0.63
honor*	-4.55	7.79	-1.89	0.58	3.84	6.99	1.78	0.55
preeminen*	-4.1	7.35	-1.81	0.56	7.23	9.75	2.41	0.74
invest	6.13	11.16	1.78	0.55	0.26	9.95	0.09	0.03

Next I look at the performance results of creating the Top 3, Top 5, and Top 10 word count variables by combining the top variables based off absolute Sharpe ratio. Table 2.3 shows the results:

Table 2.8: Top 10 Word Portfolios - Equal Weighted - LM
Top 10 equal weighted word count portfolios for the first half period and the second half period. Word portfolios are sorted by absolute Sharpe ratio value. All results are annulaized.

	Mean (%)	Std dev (%)	tstat	SR
Top 3	5.18	7.08	2.37	0.73
Top 5	5.53	6.29	2.85	0.88
Top 10	4.99	5.83	2.77	0.86

2.1 Conclusion

The study of manager sentiment in firm filings is a critical area of research that has significant implications for investors and financial markets. The approach taken in this dissertation involves a bottom-up approach that delves into the individual sentiment words used by managers and their potential impact on future stock returns. The findings of the study indicate that these individual sentiment words can be strong predictors of future stock returns, both in and out of sample.

To further explore the potential applications of these findings, the study also employs cross-sectional regression to create equal and value-weighted portfolios. The results indicate that these portfolios can be a useful tool for investors seeking to construct diversified portfolios that effectively balance risk and return. By leveraging the expected returns of individual securities and a set of risk factors, investors can achieve an optimal level of diversification across a range of asset classes and market sectors.

Moreover, the study finds that the predictive power of the word count portfolios persists over a long out-of-sample period, indicating the potential for these portfolios to generate consistent returns for investors. This is particularly noteworthy as it suggests that investors can rely on the predictive power of these portfolios when making investment decisions, thereby minimizing the risks associated with investing in individual securities.

In conclusion, the findings of this dissertation highlight the importance of considering manager sentiment in firm filings when making investment decisions. By leveraging the predictive power of individual sentiment words, investors can construct diversified portfolios that effectively balance risk and return. Furthermore, the persistent predictive power of word count portfolios suggests that these portfolios can play a significant role in diversifying risk and generating consistent returns over the long term.

References

[1] M. Baker and J. Wurgler. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680, 2006.

[2] M. Baker and J. Wurgler. Behavioral corporate finance: An updated survey. In *Handbook of the Economics of Finance*, volume 2, pages 357–424. Elsevier, 2013.

[3] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Noise trader risk in financial markets. *Journal of political Economy*, 98(4):703–738, 1990.

[4] D. GARCÍA. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

[5] R. Greenwood and A. Shleifer. Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3):714–746, 2014.

[6] D. Huang, F. Jiang, J. Tu, and G. Zhou. Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies*, 28(3):791–837, 2015.

[7] F. Jiang, J. Lee, X. Martin, and G. Zhou. Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1):126–149, 2019.

[8] T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[9] U. Malmendier and G. Tate. Ceo overconfidence and corporate investment. *The journal of finance*, 60(6):2661–2700, 2005.

[10] M. Simaan. Working with crsp/compustat in r: Reproducible empirical asset pricing. *SSRN Electronic Journal*, 2020.

[11] P. C. TETLOCK. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[12] G. Zhou. Measuring investor sentiment. *Annual Review of Financial Economics*, 10:239–259, 2018.