Textual Analysis of Annual Reports in Hedge Fund Industry

Sipeng Zeng
Shanghai Jiao Tong University
800 Dongchuan Road. Minhang District, Shanghai, 200240, China
Email: scmyzsp@sjtu.edu.cn

This Version: 30th May, 2023

Textual Analysis of Annual Reports in Hedge Fund Industry

Abstract

Combining hedge funds' web access to 10-K filings on the SEC's EDGAR server and their holdings, we show that some hedge funds who download a large number of firms' annual reports from the SEC's website adjust their positions based on the textual information in the annual reports. We find that the positions of such hedge funds are influenced by textual sentiment, textual uncertainty, and strong modal and weak modal word frequency in the annual reports. Simple long/short portfolios constructed based on textual information utilized by hedge funds can achieve an annualized alpha of over 5.2%. Taken together, our results suggest that the text in annual reports contains important information about a firm's fundamental value that is revealed by a small fraction of hedge funds that download annual reports in bulk from the SEC's website.

Keywords: Text information, Textual analysis, Hedge fund, Annual report
JEL Classification: D83, G23, G32

1. Introduction

Given their flexibility and absence of regulations, hedge fund strategies are proprietary and idiosyncratic to hedge fund managers (Chen and Liang, 2007; Kosowski et al., 2007; Agarwal et al., 2009; Sun et al., 2012; Jurek and Stafford, 2015), and the true skills are not directly observable by investors. These reasons, combined, make hedge fund selection a challenging task for investors. In general, hedge funds with private information not widely available to other investors perform better (Massound et al., 2011; Gargano et al., 2017) and those funds employ sophisticated investment strategies earn higher returns relative to their peers (Fung and Hsieh, 2000; Sun et al., 2012). More related, Crane et al. (2022) find that hedge funds could earn 1.5%-higher annualized abnormal returns than others if they actively acquire publicly available financial disclosures. This provides an empirical basis for how sophisticated hedge funds can outperform their peers by better analyzing public information and specifying trading strategies.

In terms of the type of public information obtained, Chen et al. (2020) find that funds who actively acquire companies' insider trading filings (Form 4) can have a higher return than those do not by analyzing insider trading filings. Cao et al. (2021) find that funds that actively acquire other hedge funds' position filings (13-F) can imitate the profitable positions of other funds, thereby outperforming their peers. Unlike their focus on filings related to transaction information, we focus on the use of fundamental filings by hedge funds. Specifically, we study whether hedge funds actively collect and analyze the text information in the company's annual report corresponding to the stock holdings, and adjust their positions according to the text information in the annual report.

By matching hedge fund quarterly position information with hedge fund access records to company annual report filings on SEC EDGAR, we find that hedge funds that actively access company 10-K files in bulk adjust their positions based on the textual information in the 10-K files. Using a pool of stocks whose annual reports have been downloaded in bulk by hedge funds, we find that Long-Short portfolios constructed based on textual information utilized by hedge funds can earn an annualized excess return of 3% over the last 12 years, which can be increased to 5.2% if the investment horizon is extended to 2000.

How hedge funds apply their trading strategy is often a black box, so it is almost impossible to validate whether a certain hedge fund uses text analysis methods to assist investment decisions.

However, considering that hedge funds first need to download the annual report to a local or cloud server before analyzing the text of the annual report, and then analyze the text, we adopt the precondition of text analysis as the identification strategy, that is, use crawlers and other tools to download a large number of 10-K filings. By applying Lee et al. (2015) and Cao et al.'s (2023) method for identifying machine downloads, we identify funds that use programs to batch download 10-K filings on the SEC EDGAR website. And based on this sample of machine download funds, we explore the usage of textual information in annual reports by hedge funds. We acknowledge that not all funds that have downloaded a large number of annual reports will analyze the text in the annual report, but given our conclusion that the funds that use machine-downloaded annual reports adjust their holdings based on the text information in the annual report, if we can more accurately identify those funds that use text analysis techniques, our conclusions will only be enhanced rather than weakened.

In terms of research design, we match the fund's access records on the SEC EDGAR website with the fund's holding filings. We then look at changes in the machine download funds' holdings in that company's stock during the quarter in which they downloaded the company's 10-K filing. At the same time, we take the position changes of machine download funds without downloading corresponding companies' 10-K filings as the control group. By comparing the two sets of samples, we find that the machine download funds adjust their positions according to the sentiment, the frequency of uncertain words, strong modal words, and the weak modal words of the text in the annual report they download. Take the text sentiment in the annual report as an example, if the sentiment in the annual report drops by 1 standard deviation, the machine download fund will reduce the position of the stock by 2% on average. Text uncertainty and strong and weak modal word frequency have a similar magnitude of influence on the fund's position. We do not find evidence that legal word frequency and financial constraint word frequency affect the change in fund positions. If the machine download funds do not download the annual reports of the companies corresponding to the stocks in which it holds positions, the changes in this part of the positions are not affected by the text information of the annual reports. This influence is also not found in the position changes of banks and insurance companies that download annual reports in bulk, supporting the hypothesis that hedge funds actively collect and analyze annual report text information out of a desire to gain excess returns.

To rule out the alternative hypothesis that the change in the machine downloadable fund's position is caused by an in-depth analysis of non-textual information in the annual report, we use the publication of Loughran and McDonald (2011), which significantly changes the list of negative words under financial scenarios as exogenous shocks.[1] We find that after the publication of Loughran and McDonald (2011), the textual sentiment index based on the LM dictionary began to have a significant impact on the position changes of machine download funds. Meanwhile, the negative sentiment index based on the Harvard IV-4 dictionary has a significant decrease in the impact on the position changes of machine-downloaded funds. This shows that the machine download funds actively adjust their text analysis method based on the findings of the academic community, providing evidence for hedge funds to adjust their positions based on the text information in the annual report. We also re-confirm the results using the 2018 release of Google's big language model, BERT, as an exogenous shock. The results show that immediately after the release of the 2018 BERT model, machine download funds began to adjust their positions based on the sentiment index constructed by BERT.

Subsequent analysis shows that in addition to the overall text in the annual report text, hedge funds also pay special attention to the text information in Management Discussion and Analysis (MD&A), which has an additional impact on the position changes of hedge funds. At the same time, when adjusting the positions, the cross-sectional differences in the company's text information are taken into account, and the time series differences between the text information disclosed by the company this year and last year are not considered. We also find that the impact of text sentiment on hedge fund position changes mainly comes from hedge funds selling stocks with more negative words rather than buying stocks with more positive words, which is consistent with Tetlock (2007) and Loughran and McDonald (2011)'s opinion that investors tend to focus on negative language in texts and less on positive language.

Finally, we test whether hedge funds could profit or avoid losses from trading based on textual information from annual reports. We construct long-short portfolios based on a pool of stocks whose annual reports had been downloaded by a machine download fund. We find that portfolios constructed based on sentiment, uncertainty, strong modal words, and weak modal word frequencies can achieve stable excess returns under a variety of different specifications. Over the full sample interval of 2003-

[1] About three-quarters of the negative words in the LM dictionary are different from the negative words in the Harvard IV-4 dictionary (Loughran and McDonald, 2011).

2022, all four portfolios earned an annualized excess return of about 5.2%. Portfolios constructed based on legal and financial constraint words, on the other hand, do not achieve significant excess returns, which explains the fact that hedge funds' position changes are not affected by the frequency of these two types of words. In addition to this, we discuss the difference in holdings between machine-downloaded funds and other funds and found that machine-downloaded funds are smaller than non-machine-downloaded funds. At the same time, the stocks held by machine download funds are more growth-oriented, with larger market value and weaker investment capabilities. This shows that machine download funds use their unique investment strategies to obtain excess return (Crane et al., 2022) rather than profit through more exposure to common risk factors.

Our study is related to the existing literature on public information gathering by institutional investors. In a Grossman-Stiglitz world, an agent is willing to collect information up to the private marginal value of the expected return from this activity (Grossman and Stiglitz, 1980). If the opportunity cost of paying attention to public information is too high, the performance of hedge funds will decline as they acquire more public information (Kacperczyk and Seru, 2007). Because the costs of acquiring and processing non-standardized textual information are much higher than the costs of acquiring and processing information based on standardized digits, whether hedge funds actively acquire and use textual information to formulate trading strategies is crucial to our understanding of the relationship between the costs and benefits of analyzing textual information. Our paper finds evidence that hedge funds trade based on textual information in annual reports, suggesting that in the digital age, the benefits of acquiring and analyzing textual information outweigh the costs.

Our study differs from Chen et al. (2020) and Cao et al. (2021) in two aspects. First, Chen et al. (2020) and Cao et al. (2021) acquire insider or peer trading information, which belongs to market information, and hedge funds imitate the trading of insiders or peers, which only reflects the ability of hedge funds to acquire information. In contrast, annual report information belongs to fundamental information, and how hedge funds analyze the company's public information and develop trading strategies reflects the ability of hedge funds to acquire and analyze public information. More importantly, Chen et al. (2020) and Cao et al. (2021) use structured digits while we use unstructured text, which allows us to explore the relationship between the costs and benefits of acquiring and analyzing textual data for institutional investors as described above, thus making a marginal contribution to the literature on information costs (e.g. Grossman and Stiglitz, 1980; Kacperczyk and Seru, 2007).

Our study also contributes to the exploration of the use of textual information in annual reports. Textual information in annual reports is associated with stock crash risk (Kim et al., 2019), and post-earnings announcement drift (Feldman et al., 2010), and can be used to construct investment opportunity sets (Basu et al., 2022). However, Cohen et al. (2020) show that the market response to changes in textual information in annual reports is still inadequate. The information contained in the text of annual reports is rich, but the market response to the textual information in annual reports is severely underrepresented. Do institutions in the market utilize textual information in annual reports? Our study sheds light on this question, namely that at least some hedge funds actively access and use the textual information in annual reports and can generate excess returns by analyzing the textual information. The use of textual information from annual reports in the hedge fund industry is therefore explored and validated.

More generally, our study contributes to the literature on hedge funds' performance. Hedge fund performance is linked to various fund characteristics, such as fund size, the age (Liang, 1999), managerial incentives (Ackermann et al., 1999; Liang, 1999; Edwards and Caglayan, 2001) and restrictions on hedge fund investors (Agarwal et al., 2009). In addition, hedge fund manager's skill is known to be important for generating excess return over the benchmark (e.g. Li et al., 2011; Titman and Tiu, 2011; Sun et al., 2012; Cao et al., 2013). Crane et al., (2022) discover that hedge funds actively acquire available financial disclosures. Our study then expands on the work of Crane et al. (2022) in terms of the content of public information acquired by hedge funds. That is, one of the sources of excess returns for hedge funds that actively access publicly disclosed financial information is the analysis of textual information in annual reports. Our study provides new findings to explain the sources of variation in performance among hedge funds.

2. Data and Settings

We combine data from a variety of sources to execute the empirical tests in this paper. We use CRSP to obtain stock-related information, Thomson Reuters Institutional Holdings (s34) to obtain stock holdings of funds, Compustat to obtain financial data of publicly traded companies, and I/B/E/S to obtain analyst forecast data. The 10-K text data we use is from Prof. Bill McDonald's website.[2] In this section, we describe how we construct our sample of hedge funds and how we define and identify the machine download activity of hedge funds.

2.1 Construction of the Hedge Funds Sample

As mentioned by Ben-David et al. (2013), the hedge fund list identified in the Thomson Reuters 13F database is consistent with the FactSet LionShares identification of hedge fund companies. We identify hedge funds in the Thomson Reuters 13F database as follows. Thomson Reuters database classifies institutional investors into 5 types: 1) bank trust departments, 2) insurance companies, 3) investment companies and their managers, 4) independent investment advisers, and 5) others. We first exclude institutions that are classified as type 1 or type 2.[3] Next, we manually match the remaining institutions to a list of global hedge funds provided by a third-party organization.[4] Finally, to be consistent with previous screening criteria in the hedge fund literature (e.g Jiao et al., 2016 and Cao et al., 2022), we follow Brunnermeier and Nagel (2004) and Griffin and Xu (2009) and do a final screening of the sample. For each remaining institution, we manually check its SEC ADV forms. We keep an institution if it has more than 50 percent of investment listed as "other pooled investment vehicles", including private investment companies, private equity, and hedge funds, or has more than 50 percent of clients listed as "high net worth individuals". We also require the institution to charge performance-based fees to be included in the hedge fund sample. We ended up with a total of 2,080 hedge funds from the beginning of 2000 to the mid of 2022.[5]

2.2 Matching IP Addresses with Hedge Funds

The IP addresses in the dataset are partially anonymized using a static cipher. The data describe the access of fillings by different IP addresses. A standard IP address is a combination of four numbers

[2] We sincerely thank Professor Bill McDonald for his generosity in providing data to the public. Data source: https://sraf.nd.edu

[3] It is well-known that the type classification in the 13F database is inaccurate after 1998. However, the classification errors are almost entirely driven by misclassifying type 3 or 4 institutions as type 5 institutions (Lewellen, 2011); therefore, they do not affect our sample.

[4] The third-party organization has limitations in its ability to provide a comprehensive list of hedge funds, as it can only provide data from 2007 onwards. Consequently, our sample may not include hedge funds that ceased operations before 2007, leading to potential survival bias. However, our primary conclusions are based on a sample from 2011 onwards, which are not subject to the influence of this bias.

[5] There are 1365 funds in our sample from 2000 to 2012, a number very close to the 1397 funds in Jiao et al. (2016) This gives us confidence that our screening results are reliable.

from 0 to 255, such as 123.123.123.123. But the last number of IP addresses in the dataset is replaced by a three-letter cipher, for example, 123.123.123.abc. We identify the true IP address of each observation in the log files using a look-up table (see Table 1 in Chen et al., 2020) of the cipher against the true number. Then we match organizations associated with the IP addresses to hedge funds covered by the Thomson Reuters Institutional Holdings (s34) database. Information on organizational IP addresses comes from the Whois database of the American Registry for Internet Numbers (ARIN). Because ARIN only provides slices of IP addresses from 2014 and later, if hedge funds deregistered their IP addresses before 2014, the relevant information cannot be found in ARIN. To mitigate this issue, we use another IP address book from MaxMind that can provide historical mappings of organizations to IP ranges before 2014. Finally, we can identify 678 hedge funds that could correspond to the IP addresses in the SEC EDGAR log files.

2.3 Identify Machine Download Activities of Hedge Funds

The data we use to identify machine download activity is the SEC EDGAR log file. It comprises all records of the requests for SEC fillings from EDGAR from January 2003 to June 2017.[6] Each observation in the original dataset contains information on the visitor's Internet Protocol (IP) address, timestamp, and the unique accession number of the filing that the visitor downloads.

Despite the advent of multiple data sources, the SEC EDGAR website remains the earliest and most authoritative source for company fillings to be publicly released (Cao et al., 2023). Some recent academic studies also provide evidence that investment companies rely on machine downloads of EDGAR fillings for some of their trading strategies. Crane et al. (2022) find that hedge funds that employ robotic downloads perform better than those that do not. Cao et al. (2021) show that machine downloaders exhibit skills in identifying profitable copycat trades from their peers' disclosures.

We use two criteria to measure whether an IP uses machines to download fillings, the first for IPs officially labeled by the SEC as using crawlers to access the site. The second criterion is that we identify an IP address downloading more than 50 unique firms' fillings on any given date as a machine downloader and classify all its requests in that quarter as machine downloads.[7] The criterion is the

[6] The SEC briefly suspended the availability of log files for two years and restarted them in 2020. However, starting in 2020, the log files provided by the SEC withheld the IP address and therefore could not be used in this study.

[7] We consider a situation where a hedge fund downloads fillings above the threshold (e.g., 100) on one day and then downloads fillings below the threshold (e.g., 30) on a later day. We consider such later downloaded fillings as supplements to the previously downloaded fillings. Therefore, they are also counted in samples of the machine-

same with Lee et al. (2015) and Cao et al. (2023). Figure 1 gives the logic we used to perform the classification. Ultimately, 654 hedge funds had machine download behavior in at least one quarter. There is a very clear continuity of machine downloading behavior, if a fund uses machine downloading of annual reports at year T, the probability that this fund also uses machine downloading of annual reports at year T+1 is 85.7%. In the quarters when these funds have used machine downloads, they download an average of 48% of the annual reports of the companies in their holdings. Figure 2 gives the time trend of the number of machine-download fund and the proportion of downloaded annual reports to stock holdings of these funds. The number of funds downloaded by the machine increased from 178 in 2003 to 607 in 2017, and the percentage of annual reports downloaded by these funds as a percentage of the number of stocks held grew from 30% in 2003 to 62% in 2017, both showing significant growth trends.

2.4 Construction of the Variables

Our main variables are constructed as follows. First, we calculate the percentage change in the fund's position for each quarter. Specifically, when the fund's position in stock increases relative to the previous quarter, we use equation (1) to calculate the percentage change in the position. When the fund's position in stock decreases relative to the previous quarter, we use equation (2) to calculate the percentage change in the position. In this way, we can ensure that the percentage change in the position stays from -100% to 100% when the position is first bought and -100% when it is liquidated.

$$Position\ Change_{i,j,t} = \frac{Stock\ Postion_{i,j,t} - Stock\ Postion_{i,j,t-1}}{Stock\ Postion_{i,j,t}} \tag{1}$$

$$Position\ Change_{i,j,t} = \frac{Stock\ Postion_{i,j,t} - Stock\ Postion_{i,j,t-1}}{Stock\ Postion_{i,j,t-1}} \tag{2}$$

where *i* denotes stock i, j denotes fund j, and *t* denotes quarter *t*.

We constructed six firm-year level text indices based on the LM dictionary (Loughran and McDonald, 2011; Bondnaruk et al., 2015) as our independent variables. First, we construct the first index *Sentiment* by subtracting the number of positive words from the number of negative words in the annual report and dividing it by the total number of words in the valid text of the annual report following Jiang et al., (2019). Then, we construct the second index *Uncertainty* by dividing the number of uncertain words by the total number of valid texts in the annual report. Finally, we construct the

downloaded files.

Litigation, *Model Strong*, *Model Weak*, and *Financial Constraint* indices in the same way as the *Uncertainty* index. The correlation coefficients between the indices are given in Table 1. The correlation coefficients between the indices are below 0.3, except for the correlation coefficients above 0.3 because all the words in the Model Weak dictionary are included in the Uncertainty dictionary. Meanwhile, the correlation coefficient between Model Strong and Model Weak is 0.196, showing some positive correlation. This suggests that modal words may be more of a reaction to the different ways in which words are used in the annual report rather than a difference between inevitability and likelihood.

[Insert Table 1 Here]

2.5 Regression Model Setup and Summary Statistics

We estimate the following baseline regression at a quarterly frequency to detect the use of textual information of machine download funds:

$$Position\ Change_{i,j,t} = \alpha + \beta Text\ index_{i,t} + \gamma X_{i,t} + Stock_i + Fund_j + Ind_k \times Year_t + \varepsilon_{i,j,t} \quad (3)$$

Where $Position\ Change_{i,j,t}$ refers to the percentage change in fund *j*'s position in stock *i* at quarter *t*; $Text\ index_{i,t}$ refers to text index of the annual report issued by stock *i* at quarter *t*. It includes *Sentiment*, *Uncertainty*, *Litigation*, *Model Strong*, *Model Weak*, and *Financial Constraint*. $X_{i,t}$ stacks a list of control variables. We first controlled for the stock's last quarterly return (*Return(-1)*) and volatility (*Vol(-1)*). Then referring to the Fama-French 5-factor model (Fama and French., 2016), we control company size (*Size*), book-to-market ratio (*B/M*), return on equity (*ROE*) and change in company size relative to last year (*Investment*). Finally, we control the total number of words (*Total Words*) in the annual report text. $Stock_i$ indicates stock fixed effects. $Fund_j \times Year_t$ indicates fund multiplied by year fixed effects. $Ind_k \times year_t$ indicates industry multiplied by year fixed effects. $\varepsilon_{i,j,t}$ denotes the error term. Considering the publication time of Loughran and McDonald (2011) and Bondnaruk et al. (2015), our sample covers the period 2011-2017 for regressions with *Sentiment*, *Uncertainty*, *Litigation*, *Model Strong*, and *Model Weak* as independent variables, and 2016-2017 for regressions with *Financial Constraint* as an independent variable, unless otherwise stated.[8] Summary

[8] Dictionaries for sentiment, uncertainty, litigation, strong modal words, and weak modal words were published by Loughran and McDonald (2011), while dictionaries for financially constrain words were published by Bondnaruk et al. (2015).

statistics for each variable are given in Table 2. The sample used in Panel A of Table 2 includes the position changes of all hedge funds from 2011 to 2022, while the sample used in Panel B contains the position changes of machine download funds that downloaded the firm's 10-K in that quarter. From the descriptive statistics of both samples, the text indexes and financial indicators of the stocks that 10-K has been downloaded by the machine download fund are similar to the full sample. This indicates that machine downloading funds are not selective in downloading companies' 10-K, which alleviates concerns regarding the issue of sample self-selection.

[Insert Table 2 Here]

2.6 Characteristics of machine download funds

We explore the characteristics of machine-download funds in Table 3. The sample period for this table is 2003-2022. We compared the number of shares held by machine-downloaded funds and non-machine-downloaded funds, the total amount of shares held, and the characteristics of the stocks held. Among them, the characteristic of the stock is the average value of this characteristic of all stocks held by the fund in the quarter. We find that, on average, machine download funds hold 65 more stocks than non-machine download funds, but the total value of holdings is $9.9 million lower. This indicates that the machine-download fund is relatively small and has a more diversified position. At the same time, comparing the characteristics of the stocks held by the fund, we find that the machine download fund tends to hold growth stocks, stocks with large market capitalization, and stocks with weaker investment capabilities. These characteristics are contrary to the investment strategy based on risk factors proposed by Fama and French (2015), indicating that the better investment performance of 10-K specialist funds documented by Crane et al. (2022) is not the result of having more exposure to common risk factors.

[Insert Table 3 Here]

3. Empirical Results

3.1 Baseline Regressions

In this subsection, we group the position changes of all hedge funds into three categories: Position changes of machine download funds that download the firm's 10-K during the quarter, position changes of machine download funds that do not download the firm's 10-K during the quarter, and position changes of funds that are not identified as machine download funds. Panel A of Table 4 reports the regression results for the first category of changes in positions. The results show that those funds that use machine downloads increase their holdings in companies with more positive sentiment in the text of their annual reports and decrease their holdings in companies with more uncertainty in the text of their annual reports and in companies that use more strong and weak modal words. Further decomposition of the sentiment index shows that the main reason for this result is that hedge funds sell shares of companies with a higher frequency of negative words in their annual reports, while not significantly buying shares of companies with a higher frequency of positive words in their annual reports. The results can be found in Table A2 in the Appendix. The trading of these funds is not affected by the frequency of legal and financial constraint words in the annual reports. Specifically, for every one standard deviation increase in the sentiment index (0.021) of the annual report text, the fund increased its holdings in the stock by 2.03% in the current quarter. The frequency of risky words in the annual report, strong modal words, and weak modal words are increased by one standard deviation, and the fund will reduce the position of the stock by 1.10%, 1.68%, and 1.14% respectively in the current quarter. Because not all stocks are covered by analysts and given the integrity of the trading data, we do not include earnings surprises in the control variables in the main text that would reflect the gap between actual operations and market expectations. We include standardized unexpected earnings (sue) as a control variable in Table A3, and the results remain unchanged.

The text in the annual report is indeed highly relevant to the financial situation of the company disclosed in the annual report. When a company is not doing well, the sentiment of the text in that company's annual report is also more likely to be lower than that of other companies. If a fund's position adjustment is based on financial information in the annual report related to textual information. Then, hedge funds that do not download annual reports by a machine should also be able to access and use such information, either by manually reading the annual report or pulling it from a financial data provider. Therefore, we do the same regression analysis for the second and third categories of position changes. The results are reported in Panel B and Panel C of Table 4. We could see that the regression coefficients corresponding to all text indices are not significant. The results show that if a company's annual report is not downloaded by a machine-downloaded fund, or is not identified as a machine-downloaded fund. Then the position changes are not affected by the text information.

[Insert Table 4 Here]

Next, we match the SEC log data with the bank (type 1) and insurance company (type 2) holdings data in the Thomson Reuters s34 database using the method above. In total, 145 banks and 32 insurance companies are identified as institutional investors using machine downloads. We then use regression equation (3) to investigate whether these institutional investors change their positions according to the textual information of the annual report when they download the company's 10-K filing during the quarter. The results of the regressions are presented in Table 5. The results show that banks and insurance companies do not adjust their positions based on the textual information in the annual report. This result is consistent with the reality that banks and insurance companies do not hold stocks to get excess returns from them as hedge funds do. Together, the results in Tables 4 and 5 provide valid evidence for the hypothesis that only those hedge funds that have used machines to download annual reports have transactions that are correlated with the textual information in the annual reports. This suggests that the relevant trades are based on the hedge fund's analysis of the text of the annual report rather than the financial information in the annual report.

[Insert Table 5 Here]

Last, we look at the impact of other factors on the change in the fund's position. Among the most obvious influences on changes in the fund's holdings are the stock's last quarterly return and the company's market capitalization. Both machine-downloaded and non-machine-downloaded funds tend to increase their holdings in stocks that made positive returns last quarter and in large-capitalization stocks. Funds that do not use machine downloads are also significantly affected by the level of stock volatility and company investment in the previous quarter, as they reduce their holdings in stocks with higher volatility and stocks with greater increases in company market capitalization in the previous quarter, while changes in machine downloadable funds' holdings are not significantly correlated with changes in stock volatility and company market capitalization in the previous quarter. This suggests that there is a significant difference between the investment style and risk exposure of funds that use machine downloads and those that do not.

3.2 Publication of LM dictionary: an event study approach

An intuitive alternative hypothesis is that hedge funds that bulk download 10-K files pay more attention to information in annual reports relative to other funds and analyze non-textual information in annual reports more carefully to make trading decisions. This non-textual information, in turn, is correlated with the textual information in the annual report, causing our conclusions to become unreliable. Therefore, we use the publication of Loughran and McDonald (2011) as an exogenous shock to further test the relationship between changes in hedge fund positions and textual information in annual reports.

Although relevant studies on textual analysis of annual reports existed before the publication of Loughran and McDonald (2011), the presentation of the first dictionary in finance by Loughran and McDonald (2011) is undoubtedly a landmark event, especially since the paper reconstructs to a large extent the dictionary of negative words in finance. Thus, if hedge funds do trade concerning textual information, there should be significant differences in trading patterns before and after the publication of Loughran and McDonald (2011), especially for *Sentiment* associated with negative words. With this in mind, we grouped regressions for Sentiment, Uncertainty, Model Strong, and Model Weak by year, and the regression coefficients and 95% confidence intervals are shown in Figure 3.

The findings presented in Figure 3a reveal that the impact of the Sentiment index on the trading of funds using machine downloads was not significantly significant until 2010. However, the regression coefficient of funds' trading using machine downloads on the sentiment index significantly increased from 2011 onwards, particularly in 2011 and 2012. This observation supports our earlier reasoning that the use of textual information in annual reports by hedge funds using machine downloads changed Loughran and McDonald (2011). Moving on to Figure 3b, it is evident that funds using machine downloads were already trading based on uncertainty words in annual reports before 2011. This result is in line with the findings presented by Loughran and McDonald (2016) regarding a study on textual analysis of annual reports using risk and uncertainty-related terms as early as 2003. Figures 2c and 2d give the dynamic impact of strong and weak modal words on the trading of hedge funds using machine downloads. For strong modal words, we can see that in 2011, funds using machine downloads sold a large number of stocks of companies with a high frequency of strong modal words in their annual reports, and then this relationship between strong modal words and fund position reduction gradually weakened. Because the weakly modal word dictionary is included by the uncertainty word dictionary, the dynamic impact of weakly modal word frequency on fund trading is similar to that of uncertainty words.

[Insert Figure 3 Here]

To more systematically analyze the impact of the publication of Loughran and McDonald (2011) on hedge funds' use of textual information in annual reports, we construct the following regression equation to study the effect of the publication of Loughran and McDonald (2011) on hedge fund trading:

$$Position\ Change_{i,j,t} = \alpha + \beta_1 Text\ index_{i,t} \times Machine\ download_{i,j,t} \times Post - LM_t + \beta_2 Text\ index_{i,t} \times Machine\ download_{i,j,t} + \gamma X_{i,t} + Stock_i + Fund_j + Ind_k \times Year_t + \varepsilon_{i,j,t} \quad (4)$$

where $Machine\ download_{i,j,t}$ indicates that the annual report of stock i has been downloaded by fund j using the machine in quarter t. $Post - LM_t$ is a binary variable that is 1 when the year is greater than or equal to 2011 and 0 when the year is less than 2011. The remaining interactions and control variables are denoted by $X_{i,t}$. The remaining symbols are defined in the same way as in equation (3). Only transactions from funds that have used machine downloads are included in this section. From this, the impact of the publication of Loughran and McDonald (2011) on the use of textual information by hedge funds using machine downloads can be obtained by estimating β_1. The estimation of β_2, on the other hand, yields how text information affects the trading of hedge funds that use machine downloads in the full-time period. If our hypothesis that foundations using machine downloads use downloaded annual reports for textual analysis holds, then β_1 should be significantly different from 0, especially in the regressions with sentiment words that were substantially altered by Loughran and McDonald (2011).

Although Loughran and McDonald (2011) was formally published in February 2011, the possibility exists that the article was widely distributed to potential readers before that point. Previous results also show that the effect of annual report text sentiment on changes in hedge fund positions is already significant at the 5% level of significance in 2010, suggesting that some hedge funds may have already started using the dictionary they developed at that time. However, for reasons of prudence and given that not many meetings are mentioned in the original acknowledgements, we still choose 2011 as the point in time when the event occurred.[9]

In the regression results given in Table 6, we can see that hedge funds using machine downloads are trading based on the sentiment text information in the annual report only after Loughran and

[9] Adjusting the event time point to 2010, we obtain similar results, while adjusting the event time point to 2012, the regression coefficient of $Text\ index_{i,t} \times Machine\ download_{i,j,t} \times Post - LM_t$ is no longer significant.

McDonald (2011) published. In contrast, hedge funds using machine downloads were already trading based on uncertainty word frequencies in annual reports before the publication of Loughran and McDonald (2011), and the publication of Loughran and McDonald (2011) does not significantly affect the extent to which such hedge funds use uncertainty word frequencies. The results for weak modal words are similar to those for uncertainty words. The results in column (3), on the other hand, show that the publication of Loughran and McDonald (2011) facilitated the use of the variable strong modal word frequency in the text by funds, although funds using machine downloading already showed signs of trading based on strong modal word frequency in the text before the publication of Loughran and McDonald (2011). Using the exogenous event of the publication of Loughran and McDonald's (2011) article, we find that Loughran and McDonald (2011) promote the use of sentiment words and strongly modal words in annual report texts by hedge funds that use machine downloads. These results also corroborate our hypothesis, because if the significance of the coefficients of text indices is due to financial information related to textual information in annual reports, then the publication of Loughran and McDonald (2011) should not affect the regression coefficients of fund trading on the *Sentiment*.

[Insert Table 6 Here]

To further verify our hypothesis, we also constructed a text index (*Negative_Harvard*) based on the Harvard IV-4 dictionary based on the negative word lexicon in the Harvard IV-4 dictionary. Before Loughran and McDonald (2011) substantially revised the negative vocabulary in finance, the Harvard IV-4 dictionary was the most widely used in financial text analysis at that time. If Loughran and McDonald (2011) change the basis on which hedge funds conduct text analysis, then after 2011, the negative text index based on the Harvard IV-4 dictionary should have a smaller impact on hedge fund transactions.[10]

Similar to Figure 3, Figure 4 shows the regression coefficient and 95% confidence interval of the regression by year, but the independent variable is changed to *Negative_Harvard*. From the figure, we can see that after 2011, although it is still significantly negative, the regression coefficients of *Negative_Harvard* have decreased significantly. The results of column (5) in Table 6 also support this view. The regression coefficients of the interaction term of *Negative_Harvard*, *Post-LM*, and

[10] Loughran and McDonald's negative vocabulary overlaps with Harvard IV-4's negative vocabulary by about 30%, and the impact of such common negative words on hedge fund transactions may continue after 2011. Therefore, we do not expect the negative text index based on Harvard IV-4 to completely disappear after 2011.

Downloaded are positive, which is opposite to the coefficients of the interaction of *Negative_Harvard* and *Downloaded*. It shows that after the publication of Loughran and McDonald (2011), the machine-downloaded hedge fund position changes are less affected by the frequency of negative words defined in the annual report text according to the Harvard IV-4 dictionary. This again supports the hypothesis that hedge funds analyze text in annual reports to aid trading and that Loughran and McDonald (2011) significantly change the basis for text analysis by hedge funds.

[Insert Figure 4 Here]

3.3 The Persistence of Trading Mode

Because of the limitations of the SEC EDGAR log data, our sample only covers up to the second quarter of 2017 at the latest. One might concern that hedge funds using machine downloads are no longer trading concerning textual information in annual reports, noting, in particular, the diminishing influence of textual indices on trading in such funds in Figure 3.

We notice that there is a continuum of hedge funds using machines to download annual reports. If a hedge fund downloaded its annual report by machine in year T, the probability that it will continue to download its annual report by machine in year T+1 is about 84.7%. Therefore, we collect all trades from the third quarter of 2017 to the second quarter of 2022 for funds that were identified as machine-download funds at least once from 2011 to the second quarter of 2017 to test whether hedge funds that use machine downloads have stopped trading concerning textual information in the firm's annual report. We then estimate equation (3) on this sample. If the coefficient β is significant, then it suggests that even with this more ambiguous identification, we can say that hedge funds using machine downloads consistently trade based on the textual information in the annual report.

The results of the regressions are shown in Table 7, and the coefficients of all three text indices are significant at the 5% level, except for the coefficient of modal strong, which is significant at the 10% level. This suggests that in the third quarter of 2017 and beyond, funds that have previously used machine downloads still trade based on the company's annual report text information. This trading model, based on textual information in annual reports, has not been abandoned over time.

[Insert Table 7 Here]

3.4 Out of Sample Test: The Rise of Machine Learning Methods

The rapid evolution of AI technology has profoundly changed the way computers perform text analysis. The current state-of-art natural language processing method, the Bidirectional Encoder Representation from Transformers (BERT) was introduced in 2018 by a group of researchers at Google (Devlin et al., 2018). BERT considers the sequential relation of words inside sentences and produces superior results in understanding the meaning of sentences than dictionary method. In this section, we investigate the impact of the introduction of BERT on how hedge funds use textual information in annual reports.

Because the EDGAR Log File Data Set stopped in 2017 and BERT was published in 2018, our *Machine Download* variable is not available for this test. Given that in Section 3.3 we find evidence of persistence in hedge fund behavior in the textual analysis of corporate annual reports, our sample includes hedge funds that used machine downloads in the three years prior to the publication of the BERT paper in order to roughly examine the impact of the introduction of BERT on the use of textual information in annual reports by hedge funds. Equation (5) gives the equation for the regression.

$$Position\ Change_{i,j,t} = \alpha + \beta_1 Sentiment(MD\&A)_{i,t} \times Post - BERT_t + \beta_2 Sentiment(BERT\ or\ LM)_{i,t} + \gamma X_{i,t} + Stock_i + Fund_j + Ind_k \times Year_t + \varepsilon_{i,j,t} \quad (5)$$

Unlike LM dictionary, BERT is a generic language model that has not been adapted for financial domains. Therefore, following Cao et al., (2023) we use the text from Management Discussion & Analysis (MD&A), which is closer to natural language, to compute the sentiment index based on the BERT model.[11] We also provide the results of the sentiment index based on the LM dictionary as a comparison. The BERT *Sentiment (MD&A)* is the ratio of the difference between the number of positive sentences and the number of negative sentences divided by the total number of sentences in MD&A. Considering that the BERT model was uploaded to the preprint website arXiv in 2018, we set $Post - BERT_t$ to 1 in 2019 and beyond and 0 before. In this test, the time span of our sample is 2015-2022.

The regression results are reported in Table 8. The results in column (1) show that sentiment based on the BERT model has a significant positive impact on hedge fund positions after the release of the 2018 BERT, with each standard deviation change leading to a 0.9% change in positions. This indicates

[11] We train the BERT model using the GoEmotions dataset provided by Google, a high quality dataset containing more than 58,000 manually annotated Reddit comments and widely used in the training of machine learning models (Demszky et al., 2020). Using different datasets to train the BERT model will give slightly different results.

that after the release of BERT, Machine Download Fund actively started to use this new technology to assist their analysis of annual report texts. The negative but insignificant coefficient on the interaction term in column (2) indicates that the reliance of the machine download fund on LM dictionary does not significantly diminish after the release of BERT. This may be because the LM dictionary is designed for text analysis in the financial domain and therefore recognizes the emotions that are exclusively found in financial texts.[12] Overall, we find evidence of active adoption of new text analytics techniques by hedge funds.

[Insert Table 8 Here]

4. Additional Analysis

4.1 Sub-sectional Analysis

Some information of interest to investors is reported in the corresponding sub-sections of the Annual Report. For example, Item 1A - "Risk Factors" includes information about the most significant risks that apply to the company or to its securities. This sub-section disclosing the company's business risks is often of high interest to investors (Campbell et al., 2013; Hope et al., 2016). Item 7 - "Management's Discussion and Analysis of Financial Condition and Results of Operations" gives the company's perspective on the business results of the past financial year. This section, known as the MD&A for short, allows company management to tell its story in its own words. Therefore, it has also received extensive attention from the literature related to text analysis and investors (eg. Hoberg and Lewis, 2017; Murphy et al., 2018; Lo et al., 2017).

In this subsection, we analyze whether machine download funds pay special attention to these two subsections in the annual report. We use the same approach to construct text indices for the subsections Item 1A - "Risk Factors" and Item 7 - "MD&A" and use equation (6) to estimate the net impact of the subsections' text indices on fund trades using machine downloads.

$$Percent\ Change_{i,j,t} = \alpha + \beta_1 Text\ index\ (subsection)_{i,t} + \beta_2 Text\ index_{i,t} + \gamma X_{i,t} + Stock_i + Fund_j + Ind_k \times Year_t + \varepsilon_{i,j,t} \quad (6)$$

[12] In 2022, Huang et al. (2022) released a modified version of the BERT model based on financial domains, the FinBERT. However, the model was published too far after 2017, when SEC EDGAR stopped providing log files, to be included in our analysis.

Where $Text\ index\ (subsection)_{i,t}$ indicates the text index of the corresponding subsection of stock i at quarter t. The remaining symbols are defined in the same way as in equation (3). β_1 denotes the net impact of the subsection text index on hedge fund trading. The sample period for each regression is the same as in section 3.1.

The regression results are shown in Table 9. Panel B of Table 9 shows that the regression coefficient for *Litigation(Risk)* is positive and significant at the 10% level, suggesting that some machine download funds may be concerned about the legal situation in the risk-disclosure section and buy stocks of companies that mention legal topics more in their annual report Item-1A. The regression coefficients of the remaining text indices of the risk-disclosure section are all insignificant.

The results in column (1) of Table 9, Panel B, show that the coefficient of *Sentiment(MD&A)* is positive and significant at the 1% level, while the coefficient of *Sentiment(Full text)* is no longer significant. This suggests that hedge funds' focus on textual sentiment in annual reports is primarily a focus on textual sentiment in MD&A. This finding is consistent with the conclusion mentioned in Loughran and McDonald (2016) that because management is relatively free to write MD&A subsections, MD&A is a better indicator of management's sentiment about the company's business conditions than the full annual report. The regression coefficients of *Uncertainty(MD&A)*, *Modal Strong(MD&A)*, and *Modal Weak(MD&A)* all have the same negative signs as the full-text index and are significant at the 10% or 5% level. This indicates that hedge funds pay additional attention to the text of the MD&A subsection in addition to the full text of the annual report.

[Insert Table 9 Here]

4.2 Which Matters, Cross-sectional Variances or Time-series Variances?

All our regressions in the previous subsection are based on the level of variables and do not take into account the variation in the time series of variables. In this subsection, we explore whether hedge funds take into account time-series changes in the firm text index when trading.

We use equation (7) to calculate the difference in the text index in the company's annual report for the current year relative to the previous year.[13] Then we added $\Delta Text\ index_{i,t}$ to the regression equation (3).

[13] To facilitate the presentation in the table, we scaled the calculated $\Delta Text\ index_{i,t}$ value by 100.

$$\Delta Text\ index_{i,t} = \frac{Text\ index_{i,t} - Text\ index_{i,t-1}}{Text\ index_{i,t-1} \times 100} \tag{7}$$

The results of the regressions are shown in Table 10, and the regression coefficients of all four $\Delta Text\ index_{i,t}$ are not significant, indicating that hedge funds do not consider changes in text information relative to last year when using text information in annual reports, but only consider cross-sectional variances in text information in annual reports to adjust their positions.

[Insert Table 10 Here]

4.3 Does Text Information Predict Return?

Crane et al., (2022) show that trading in stocks whose 10-K documents are actively acquired by hedge funds predicts abnormal stock returns.[14] We explore whether the textual information in the annual report is related to such abnormal stock returns. Specifically, concerning our previous results, we constructed six portfolios based on stocks whose annual reports were machine-downloaded during the disclosure quarter:[15] 1) long stocks with *Sentiment* above the median and short stocks with *Sentiment* below the median; 2) long stocks with *Uncertainty* below the median and short stocks with *Uncertainty* above the median; 3) long stocks with *Litigation* below the median and short *Litigation* above the median; 4) long stocks with *Modal Strong* below the median and short stocks with *Modal Strong* above the median; 5) long stocks with *Modal Weak* below the median and short *Modal Weak* above the median; 6) long stocks with *Financial Constrain* below the median and short stocks with *Financial Constrain* above the median. All stocks are weighted by market value and the positions are switched on June 30 of each year in the same way as Fama and French (1993, 2016).

Table 11 reports the returns of the portfolio. Portfolio performance is measured by the mean monthly excess return over the risk-free rate, and the risk-adjusted return using CAPM, the Fama-French three-factor model (Fama and French, 1993), and the Carhart four-factor model (Carhart, 1997).[16] Panel A of Table 11 reports the α obtained for the constructed portfolio from June 2003 to December 2022 for the returns under different models, and the portfolios constructed based on *Sentiment*, *Uncertainty*, *Modal Strong*, and *Modal Weak* obtain significant α at the 1% level under all

[14] For details, see Table 7 in Crane et al. (2022).

[15] The specific screening criteria is that if the annual report of stock A has been downloaded by at least one machine download fund in the quarter when it is released, then the stock will be selected into our stock pool.

[16] The returns of factors are from WRDS' Fama-French factor return database.

four models. All four portfolios earned a monthly excess return of about 42 basis points, around 5.2% annualized. A portfolio constructed based on Litigation exhibits insignificant α, and a portfolio constructed based on *Financial Constrain* yields a monthly excess return of about 31 basis points, but only significant at 10% level. This suggests that in general the textual information in the annual report can predict the future returns of the stock. And hedge funds correctly identify and utilize the useful part of the information in the text of the annual report.

Next, we explore whether the portfolio excess returns are affected after the release of the LM dictionary (Loughran and McDonald, 2011; Bondnaruk et al., 2015). We restrict the sample interval to the period after the publication of the paper to December 2022 and then redo the regressions. Panel B of Table 11 shows the results of the regressions. Panel B of Table 11 presents the results of the regressions, and the results in column (1) show that the portfolio constructed based on Sentiment can also obtain significant excess returns at the 10% level under the Fama-French 3-factor model, while the excess returns disappear when the momentum factor is added. This suggests that after the publication of Loughran and McDonald (2011), investors widely focus on and utilize the sentiment information in annual reports, the excess returns thus significantly weakened or even disappear. The results in columns (2)(4)(5) show that the portfolios based on *Uncertainty*, *Modal Strong*, and *Modal Weak* can still obtain a monthly excess return of 25 basis points after the publication of the paper, which is equivalent to 3% annualized. The excess returns of the portfolios constructed based on *Financial Constraints* are also no longer significant after the publication of the paper. These results indicate that hedge funds can correctly identify textual information in annual reports and can achieve excess returns from trading based on textual information. These results also show that the excess returns obtained by portfolios constructed based on textual information significantly decrease or even disappear after the publication of the relevant papers, suggesting that hedge funds also keep an eye on the academic results and adjust their investment strategies based on the incremental information provided by the papers.

[Insert Table 11 Here]

5． Concluding Remarks

We have discovered that hedge funds possessing substantial access to 10-K information engage in systematic trading activities based on the textual information contained within annual reports. These trades can generate excess returns for such hedge funds. Specifically, funds that bulk downloaded annual reports increase their holdings in stocks with more positive textual sentiment in the annual report and decrease their holdings in stocks with more uncertainty words and more strong and weak modal words in the annual report. We do not find evidence that hedge funds trade based on legal word frequencies and financial constraint word frequencies in annual reports.

Further analysis shows that hedge funds also specifically analyze textual information in the Management Discussion and Analysis section of annual reports and that hedge funds' usage of textual information is influenced by academic papers. The results of our study provide evidence that hedge funds are utilizing text analytics capabilities and associated investment strategies to generate profits. Moreover, our research serves as a supplementary and extensive contribution to the work of Chen et al. (2017), as it sheds light on the substantial differences among hedge funds in terms of their level of expertise and the investment strategies they employ. This is accomplished through our empirical validation of the integration of textual data by hedge funds into their investment decision-making processes.

Reference

Ackermann, C., McEnally, R., & Ravenscraft, D. (1999). The performance of hedge funds: Risk, return, and incentives. *The Journal of Finance*, 54(3), 833-874.

Agarwal, V., Daniel, N. D., & Naik, N. Y. (2009). Role of managerial incentives and discretion in hedge fund performance. *The Journal of Finance*, 64(5), 2221-2256.

Basu, S., Ma, X., & Briscoe-Tran, H. (2022). Measuring multidimensional investment opportunity sets with 10-K text. *The Accounting Review*, 97(1), 51-73.

Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly Journal of Economics*, 128(4), 1547-1584.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4), 623-646.

Brunnermeier, M., & Nagel, S. (2004). Hedge funds and the technology bubble. *The Journal of Finance*, 59(5), 2013-2040.

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H. M., & Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19, 396-455.

Cao, C., Chen, Y., Liang, B., & Lo, A. W. (2013). Can hedge funds time market liquidity?. *Journal of Financial Economics*, 109(2), 493-516.

Cao, S., Da, Z., Jiang, D., & Yang, B. (2022). Do Hedge Funds Strategically Misreport Their Holdings? Evidence from 13F Restatements. Working Paper.

Cao, S., Du, K., Yang, B., & Zhang, A. L. (2021). Copycat skills and disclosure costs: Evidence from peer companies' digital footprints. *Journal of Accounting Research*, 59(4), 1261-1302.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023). How to talk when a machine is listening?: Corporate disclosure in the age of AI. *Review of Financial Studies*, Forthcoming.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 57-82.

Chen, H., Cohen, L., Gurun, U., Lou, D., & Malloy, C. (2020). IQ from IP: Simplifying Search in portfolio choice. *Journal of Financial Economics*, 138(1), 118-137.

Chen, Y., Cliff, M., & Zhao, H. (2017). Hedge funds: The good, the bad, and the lucky. *Journal of Financial and Quantitative Analysis*, 52(3), 1081-1109.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, 75(3), 1371-1415.

Crane, A., Crotty, K., & Umar, T. (2022). Hedge funds and public information acquisition. *Management Science*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Edwards, F. R., & Caglayan, M. O. (2001). Hedge fund performance and manager skill. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 21(11), 1003-1028.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3-56.

Fama, E. F., & French, K. R. (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies*, 29(1), 69-103.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift, and accruals. *Review of Accounting Studies*, 15, 915-953.

Fung, W., & Hsieh, D. A. (2000). Performance characteristics of hedge funds and commodity funds: Natural vs. spurious biases. *Journal of Financial and Quantitative Analysis*, 35(3), 291-307.

Gargano, A., Rossi, A. G., & Wermers, R. (2017). The Freedom of information act and the race toward information acquisition. *The Review of Financial Studies*, 30(6), 2179-2228.

Grossman, S. J., & Stiglitz, J. E. (1980). Stockholder unanimity in making production and financial decisions. *The Quarterly Journal of Economics*, 94(3), 543-566.

Griffin, J. M., & Xu, J. (2009). How smart are the smart guys? A unique view from hedge fund stock holdings. *The Review of Financial Studies*, 22(7), 2531-2570.

Hoberg, G., & Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure?. *Journal of Corporate Finance*, 43, 58-85.

Hope, O. K., Hu, D., & Lu, H. (2016). The benefits of specific risk-factor disclosures. *Review of Accounting Studies*, 21, 1005-1045.

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40, 806-841

Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1), 126-149.

Jiao, Y., Massa, M., & Zhang, H. (2016). Short selling meets hedge fund 13F: An anatomy of informed demand. *Journal of Financial Economics*, 122(3), 544-567.

Jurek, J. W., & Stafford, E. (2015). The cost of capital for alternative investments. *The Journal of Finance*, 70(5), 2185-2226.

Kacperczyk, M., & Seru, A. (2007). Fund manager use of public information: New evidence on managerial skills. *The Journal of Finance*, 62(2), 485-528.

Kim, C., Wang, K., & Zhang, L. (2019). Readability of 10-K reports and stock price crash risk. *Contemporary Accounting Research*, 36(2), 1184-1216.

Kosowski, R., Naik, N. Y., & Teo, M. (2007). Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. *Journal of Financial Economics*, 84(1), 229-264.

Li, H., Zhang, X., & Zhao, R. (2011). Investing in talents: Manager characteristics and hedge fund performances. *Journal of Financial and Quantitative Analysis*, 46(1), 59-82.

Liang, B. (1999). The performance of hedge funds: Risk, return, and incentives. *Financial Analysts Journal*, 55(4):72–85.

Lee, C. M., Ma, P., & Wang, C. C. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, 116(2), 410-431.

Lewellen, J. (2011). Institutional investors and the limits of arbitrage. *Journal of Financial Economics*, 102(1), 62-80.

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, 63(1), 1-25.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230.

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, 12, 357-375.

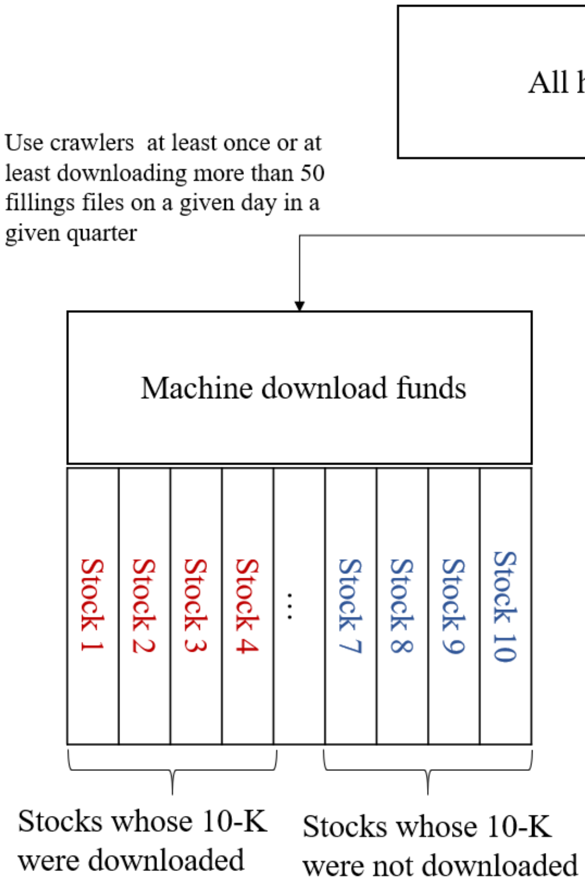
Massoud, N., Nandy, D., Saunders, A., & Song, K. (2011). Do hedge funds trade on private information? Evidence from syndicated lending and short-selling. *Journal of Financial Economics*, 99(3), 477-499.

Murphy, P. R., Purda, L., & Skillicorn, D. (2018). Can fraudulent cues Be transmitted by innocent participants?. *Journal of Behavioral Finance*, 19(1), 1-15.

Sun, Z., Wang, A., & Zheng, L. (2012). The road less traveled: Strategy distinctiveness and hedge fund performance. *The Review of Financial Studies*, 25(1), 96-143.

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

Titman, S., & Tiu, C. (2011). Do the best hedge funds hedge?. *The Review of Financial Studies*, 24(1), 123-168.


All hedge funds
Use crawlers at least once or at least downloading more than 50 fillings files on a given day in a given quarter
Else
Machine download funds
Non-machine download funds
Stock 1
Stock 2
Stock 3
Stock 4
…
Stock 7
Stock 8
Stock 9
Stock 10
Stocks whose 10-K were downloaded
Stocks whose 10-K were not downloaded

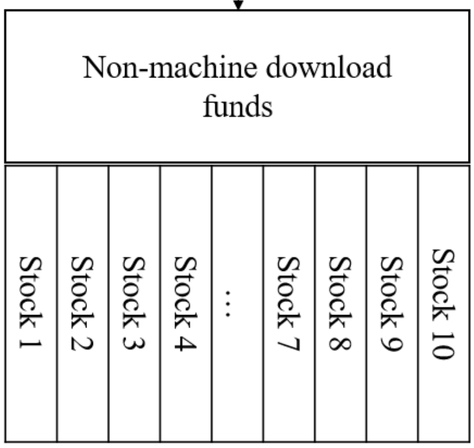


Figure 1 Flowchart of fund and position classification

This flowchart reports how hedge funds and hedge fund positions are categorized. If a hedge fund has used a crawler to download a file in a quarter or has downloaded more than 50 fillings in a day, then that hedge fund is classified as a machine download fund for that quarter. Conversely, it is classified as a non-machine download fund. If the 10-K of stock in the machine-downloadable fund's holdings is downloaded by the fund, the change in the fund's position in that stock for that quarter will be classified as a transaction based on the 10-K. If the 10-K of stock in the machine-downloadable fund's holdings is not downloaded by the fund, the change in the fund's position in that stock for that quarter will be classified as a transaction not based on 10-K.

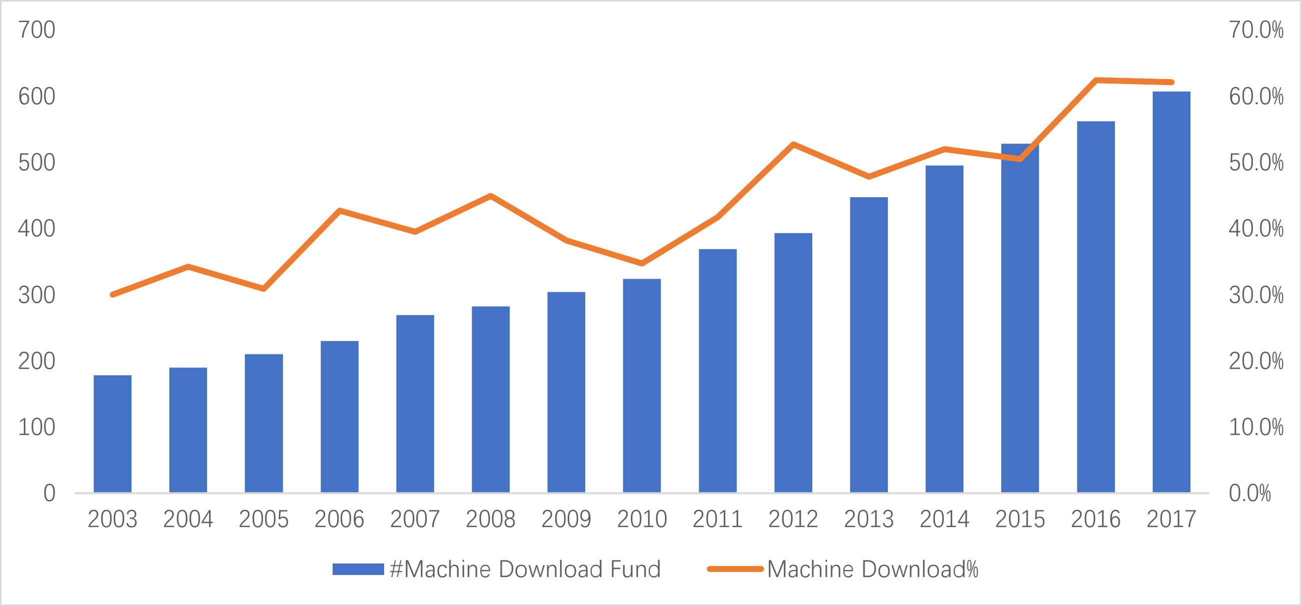


Figure 2 Time trends of the number of hedge funds using machine downloads and the percentage of stocks whose annual reports are downloaded in the positions of machine-download funds.

This figure shows the number of funds using machine downloads, and the number of positions whose annual reports were downloaded using the machine as a percentage of the total number of positions in machine download funds from 2003 to 2017. The blue bar in the figure gives the number of funds downloaded by machine each year, with the axis on the left. The orange line in the figure is the percentage of stocks that had their annual reports downloaded by machine-download funds as a percentage of the total number of stocks held by machine-downloaded funds for each year, with the axis on the right.

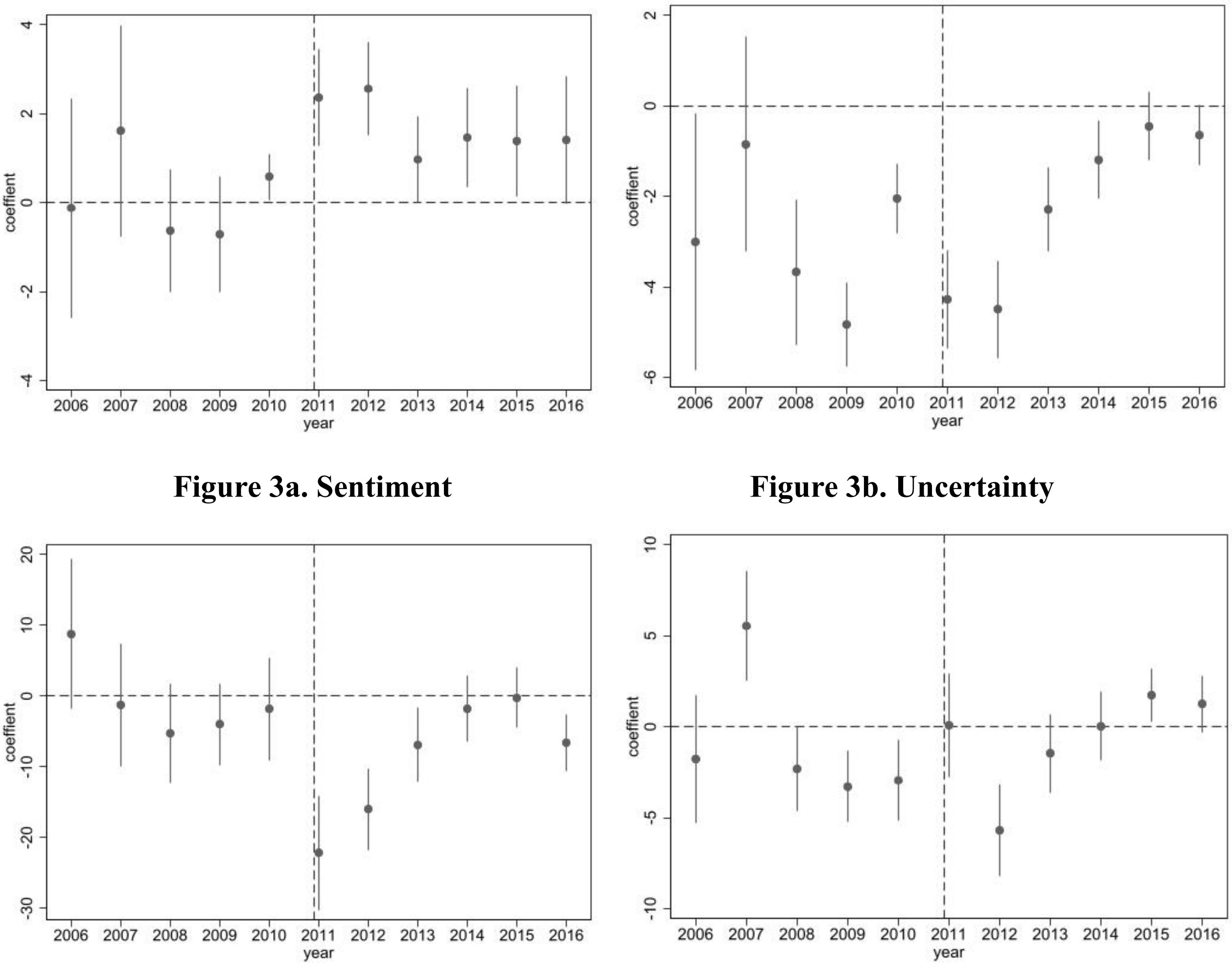


Figure 3a. Sentiment

Figure 3b. Uncertainty

Figure 3c. Model Strong

Figure 3d. Model Weak

Figure 3. The Dynamic Position Change of Machine Download Funds Based on LM Textual Indices

This figure gives the regression coefficients of the change in machine download fund positions on the text index from 2006 to 2016. The independent variable used in Figure 3a is *Sentiment*, the independent variable used in Figure 3b is *Uncertainty*, the independent variable used in Figure 3c is *Model Strong* and the independent variable used in Figure 3d is *Modal Weak*. The dependent variable is *Position Change*. The regression equation is set according to equation (3). For variable definitions, see Table A1. The vertical dashed line in the figure indicates the time of publication of Loughran and McDonald (2011). The whiskers in the graph indicate the 95% confidence interval.

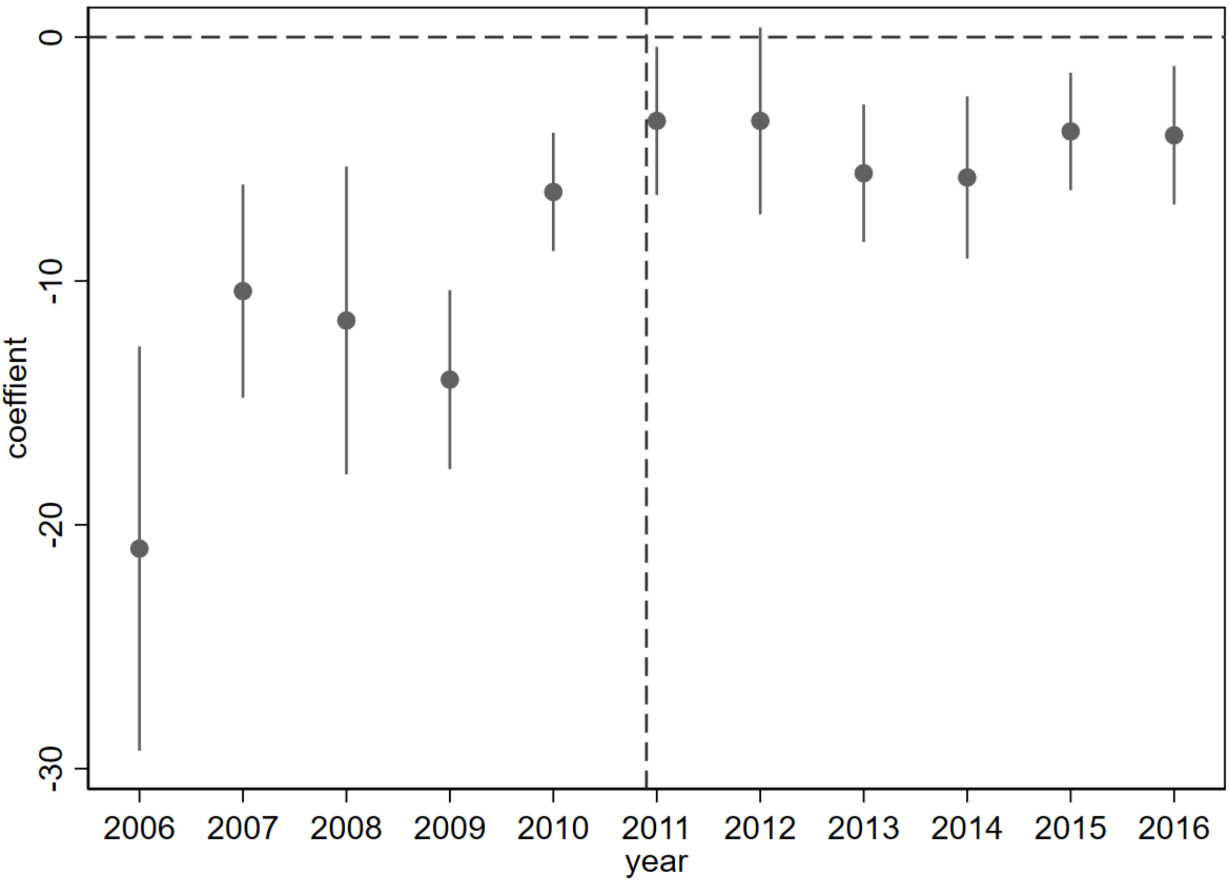


Figure 4. The Dynamic Position Change of Machine Download Funds Based on Harvard IV-4 Dictionary Negative Words

This figure gives the regression coefficients of the change in machine download fund positions on the Harvard IV-4 Negative index from 2006 to 2016. The independent variable is *Negative_Harvard.* The dependent variable is *Position Change*. The regression equation is set according to equation (3). For variable definitions, see Table A1. The vertical dashed line in the figure indicates the time of publication of Loughran and McDonald (2011). The whiskers in the graph indicate the 95% confidence interval.

Table 1 Correlation of Text Indexes

This table gives the correlation coefficients between two pairs of each text index. The definitions of variables are given in Table A1.

	Sentiment	Uncertainty	Litigation	Modal Strong	Modal Weak	Fin Constraint
Sentiment	1					
Uncertainty	-0.193	1				
Litigation	-0.122	-0.022	1			
Modal Strong	0.025	0.193	-0.089	1		
Modal Weak	-0.160	0.674	0.131	0.196	1	
Fin Constraint	-0.115	0.083	0.091	0.061	0.087	1

Table 2 Summary Statistics

This table gives summary statistics. The sample included in Panel A is all hedge fund position changes between 2011 and 2017. The sample included in Panel B is position changes based on downloaded 10-K by machine download hedge funds. Variables are defined in Table A1.

Panel A: All hedge fund position changes during 2011-2017

Variables	Obs	Mean	Std.Dev	P1	Median	P99
Dependent Variable						
Position Change	2,543,210	-0.114	0.474	-1	0	1
Independent Variables						
Sentiment	2,543,210	-0.015	0.007	-0.032	-0.015	0.003
Uncertainty	2,543,210	0.021	0.005	0.005	0.021	0.032
Litigation	2,543,210	0.012	0.005	0.004	0.011	0.028
Modal Strong	2,543,210	0.003	0.001	0.0003	0.003	0.006
Modal Weak	2,543,210	0.013	0.004	0.002	0.013	0.021
Fin Constraint	2,543,210	0.008	0.003	0.002	0.008	0.015
Control Variables						
Return(-1)	2,543,210	0.056	0.184	-0.433	0.049	0.569
Vol(-1)	2,543,210	0.065	0.052	0.004	0.053	0.240
Size	2,543,210	8.211	1.819	3.995	8.253	12.309
B/M	2,543,210	0.542	0.484	0.027	0.425	2.104
ROE	2,543,210	0.113	1.117	-1.213	0.109	1.291
Investment	2,543,210	0.116	0.278	-0.306	0.054	1.704
Total words	2,543,210	9.575	0.658	7.288	9.619	10.905
Sue	1,967,664	0.940	9.757	-11.572	0.674	15.106

Panel B: Position changes based on downloaded 10-K during 2011-2017

Variables	Obs	Mean	Std.Dev	P1	Median	P99
Dependent Variable						
Position Change	550,799	-0.102	0.481	-1	0	1
Independent Variables						
Sentiment	550,799	-0.015	0.007	-0.032	-0.015	0.003
Uncertainty	550,799	0.021	0.005	0.005	0.021	0.032
Litigation	550,799	0.012	0.005	0.004	0.012	0.028
Modal Strong	550,799	0.003	0.001	0.0003	0.003	0.006
Modal Weak	550,799	0.013	0.004	0.002	0.013	0.021
Fin Constraint	550,799	0.008	0.003	0.002	0.008	0.015
Control Variables						
Return(-1)	550,799	0.057	0.185	-0.439	0.052	0.572
Vol(-1)	550,799	0.068	0.053	0.004	0.055	0.245
Size	550,799	8.162	1.850	3.912	8.201	12.316
B/M	550,799	0.549	0.503	0.027	0.430	2.149
ROE	550,799	0.099	1.120	-1.384	0.106	1.291
Investment	550,799	0.109	0.268	-0.306	0.052	1.626
Total words	550,799	9.556	0.662	7.231	9.602	10.889
Sue	498,929	0.913	9.059	-11.490	0.707	14.142

Table 3 Differences in characteristics between machine-download funds and non-machine-download funds

This table gives the characteristics of machine download and non-machine download fund holdings. The sample unit used in this table is fund-quarter, and all stock-level metrics are averages of metrics for stocks held by the fund in a given quarter. The sample period covers a total of 80 quarters over 20 years from 2003 to 2022. Machine download funds are defined as hedge funds that applied machine downloads in at least one quarter from the first quarter of 2011 through the second quarter of 2017. The Diff column uses a t-test to test the difference between the means of the two groups of funds. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

	Machine Download Fund		Non-machine Download Fund		
	Mean	Std.dev	Mean	Std.dev	Diff
Number of stock holdings	338.676	209.89	274.156	167.85	64.520***
Total value of stock holdings ($m)	29.096	57.419	38.964	106.45	-9.868***
Return(-1)	0.074	2.448	0.054	1.904	0.020*
Vol(-1)	0.065	0.054	0.066	0.057	0.001
B/M	0.432	0.424	0.443	0.436	-0.011***
Size	8.548	1.944	8.466	1.992	0.082***
Roe	0.162	0.439	0.159	0.905	0.003
Inv	0.113	0.249	0.119	0.259	-0.005***
N of fund-quarter	22,015		49,445		

Table 4 Hedge Fund Trading Based on Textual

This table reports regressions of quarterly Position Change of positions on different text indices by position types. The unit of observation is a hedge fund-quarter-stock holding. The dependent variable for all columns is *Position Change*. The sample in Panel A contains all positions of machine download funds that have downloaded the companies' 10-K in that quarter. The sample in Panel B contains all positions in companies' 10-K that the machine download fund has not downloaded in that quarter. The sample in Panel C contains the positions of all non-machine downloadable funds. The dictionaries for each category of words are from Loughran and McDonald (2011) and Loughran and McDonald (2015). All regressions contain industry-year, stock, and hedge fund-year fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

Panel A: Transactions based on 10-K downloaded by machine-download funds

	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment	0.966**					
	(0.336)					
Uncertainty		-1.573**				
		(0.489)				
Litigation			0.241			
			(0.585)			
Modal Strong				-6.734**		
				(2.591)		
Modal Weak					-1.900**	
					(0.531)	
Fin Constraint						0.480
						(1.046)
Return(-1)	0.036***	0.036***	0.036***	0.037***	0.036***	0.036***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Vol(-1)	-0.007	-0.009	-0.008	-0.008	-0.009	-0.008
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
Size	0.059***	0.059***	0.059***	0.059***	0.059***	0.059***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
B/M	-0.012*	-0.012*	-0.012*	-0.012*	-0.012*	-0.012*
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
ROE	0.002	-0.002	-0.002	-0.002	-0.002	-0.002
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Investment	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Total words	-0.007*	-0.006*	-0.008**	-0.007*	-0.006*	-0.007*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	540,213	540,213	540,213	540,213	540,213	347,307
Adj R^2	0.128	0.128	0.128	0.128	0.128	0.159

Panel B: Transactions based on 10-K that have not been downloaded by machine-download funds

	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment	-0.579					
	(0.396)					
Uncertainty		0.175				
		(0.571)				
Litigation			0.209			
			(0.593)			
Modal Strong				-1.755		
				(1.733)		
Modal Weak					-1.257	
					(0.657)	
Fin Constraint						1.786
						(1.793)
Return(-1)	0.045***	0.045***	0.045***	0.045***	0.045***	0.047**
	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.015)
Vol(-1)	-0.036	-0.036	-0.035	-0.035	-0.035	-0.008
	(0.029)	(0.031)	(0.030)	(0.029)	(0.030)	(0.038)
Size	0.031***	0.029**	0.031***	0.030***	0.031***	0.006
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)
B/M	-0.002	-0.002	-0.002	-0.002	-0.002	0.006
	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)	(0.038)
ROE	0.002	-0.004	-0.002	-0.002	-0.003	-0.003
	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
Investment	-0.008*	-0.007*	-0.007*	-0.008*	-0.008*	0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.002)
Total words	-0.006*	-0.004*	-0.005	-0.004	-0.006*	-0.012*
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	450,517	450,517	450,517	450,517	450,517	246,625
Adj R^2	0.081	0.081	0.081	0.081	0.081	0.138

Panel C: Transaction by funds that do not use machine downloads

	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment	-0.580					
	(0.404)					
Uncertainty		0.554				
		(0.297)				
Litigation			0.717			
			(0.525)			
Modal Strong				-2.306		
				(1.721)		
Modal Weak					0.309	
					(0.839)	
Fin Constraint						0.645
						(0.767)
Return(-1)	0.054***	0.054***	0.055***	0.055***	0.054***	0.055***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Vol(-1)	-0.053**	-0.052**	-0.053**	-0.052**	-0.052**	-0.052**
	(0.017)	(0.016)	(0.017)	(0.017)	(0.017)	(0.017)
Size	0.024***	0.024***	0.024***	0.024***	0.024***	0.024***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
B/M	-0.002	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
ROE	0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Investment	-0.014**	-0.014**	-0.014**	-0.014**	-0.0014**	-0.0014**
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Total words	-0.003	-0.003	-0.001	-0.001	-0.003	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	4,303,890	4,303,890	4,303,890	4,303,890	4,303,890	2,755,809
Adj R^2	0.080	0.080	0.080	0.080	0.080	0.129

Table 5 Banks and Insurance Companies Trading Based on Textual

This table presents the results of regression analyses that examine the relationship between textual indices of stocks and quarterly Position Change of positions among banks and insurance companies. The sample consists of positions held by banks and insurance companies whose 10-K reports were downloaded during the quarter. Banks and insurance companies are identified by type 1 and type 2 of the institutional category variable in the Thomson Reuters s34 database. The dependent variable for all columns is *Position Change*. The dictionaries for each category of words are from Loughran and McDonald (2011) and Loughran and McDonald (2015). All regressions contain industry-year, stock, and institute-year fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment	0.255					
	(0.381)					
Uncertainty		0.108				
		(0.453)				
Litigation			0.316			
			(0.757)			
Modal Strong				-4.018		
				(2.944)		
Modal Weak					0.242	
					(0.623)	
Fin Constraint						-0.137
						(0.453)
Return(-1)	0.018	0.018	0.018	0.018	0.018	0.018
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.019)
Vol(-1)	0.033	0.032	0.032	0.032	0.032	0.058
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
Size	0.053**	0.053**	0.053**	0.053**	0.053**	-0.007
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.013)
B/M	-0.029**	-0.029**	-0.029**	-0.029**	-0.029**	-0.028*
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.013)
ROE	0.001	0.001	0.001	0.001	0.001	0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Investment	0.006	-0.006	-0.006	-0.006	-0.006	0.004
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.003)
Total words	-0.006*	-0.007*	-0.006*	-0.006*	-0.006*	-0.008
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Institute FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	429,875	429,875	429,875	429,875	429,875	316,034
Adj R^2	0.092	0.092	0.092	0.092	0.092	0.104

Table 6 Hedge Fund Trading Based on Textual: Event Study Approach

This table uses a triple difference approach to examine the effect of the publication of Loughran and McDonald (2011) on the use of textual information by hedge funds for trading. The sample used in this table includes all position changes of the machine-downloaded funds, regardless of whether the company's 10-K has been downloaded or not. The *Index* variable in the four columns is listed as follows: *Sentiment* in the first column, *Uncertainty* in the second column, *Modal Strong* in the third column, *Modal Weak* in the fourth column, and *Negative_Harvard* in the fifth column. *Post-LM* is equal to 1 when the year is greater than or equal to 2011, and 0 otherwise. *Downloaded* equals 1 when the company's 10-K has been downloaded by the fund in the quarter, otherwise, it is 0. The coefficient of *Index×Post-LM×Downloaded* captures the effect of the text index on the change in the fund's position after the publication of Loughran and McDonald (2011). The coefficient of *Index×Downloaded* captures the effect of the text index on the change in the fund's position before the publication of Loughran and McDonald (2011). The remaining interaction terms and individual variables are also added to the control variables. The dictionary for each category of words is from Loughran and McDonald (2011). All regressions contain industry-year, stock, and hedge fund-year fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

	(1)	(2)	(3)	(4)	(5)
Dep variable:			Position Change		
Index:	Sentiment	Uncertainty	Modal Strong	Modal Weak	Negative_Harvard
Index×Post-LM×	2.944***	1.720	-9.117*	2.261	5.534*
Downloaded	(0.762)	(1.086)	(4.519)	(1.529)	(3.062)
Index×Downloaded	0.218	-4.166***	-5.309*	-6.117***	-12.783***
	(0.679)	(0.981)	(4.159)	(1.393)	(3.212)
Index×Post-LM	-0.616	-2.016*	-0.334	-2.065*	1.306
	(0.585)	(1.101)	(3.592)	(1.118)	(2.540)
Post-LM×Downloaded	0.068***	0.026	0.071***	0.025	0.040**
	(0.012)	(0.027)	(0.017)	(0.024)	(0.018)
Index	0.108	2.909	1.458	2.825	2.480
	(0.511)	(1.777)	(3.662)	(1.915)	(2.393)
Downloaded	-0.060***	0.015	-0.044	-0.001	-0.035**
	(0.008)	(0.022)	(0.013)	(0.018)	(0.013)
Controls	Yes	Yes	Yes	Yes	Yes
Year×Industry FE	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes
Obs	1,147,505	1,147,505	1,147,505	1,147,505	1,147,505
Adj R^2	0.084	0.083	0.084	0.075	0.075

Table 7 The Persistence of Trading Mode

This table examines the impact of annual report text indexes on machine-download funds' position changes after the third quarter of 2017. The sample employed in this tabulation comprises all instances of position changes that occurred between the third quarter of 2017 and the second quarter of 2022 for hedge funds that applied machine downloads in at least one quarter from the first quarter of 2011 through the second quarter of 2017. The dependent variable for all columns is *Position Change*. The dictionary for each category of words is from Loughran and McDonald (2011). All regressions contain industry-year, stock, and hedge fund-year fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

	(1)	(2)	(3)	(4)
Sentiment	0.917**			
	(0.239)			
Uncertainty		-0.884**		
		(0.271)		
Modal Strong			-6.139*	
			(2.470)	
Modal Weak				-1.091**
				(0.329)
Controls	Yes	Yes	Yes	Yes
Year×Industry FE	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes
Obs	804,871	804,871	804,871	804,871
Adj R^2	0.076	0.076	0.076	0.076

Table 8 Hedge Fund Adoption of New AI Technology

This table gives the impact of the MD&A sentiment index constructed based on two different methods, BERT and LM dictionary, on the change in hedge fund positions before and after the release of BERT. *Sentiment (MD&A)* in column (1) is the difference between positive and negative sentences in the MD&A section determined by the BERT model divided by the total number of sentences. The *Sentiment (MD&A)* in column (2) is the difference between the number of positive words and the number of negative words in the MD&A chapter determined by the LM dictionary divided by the total number of words. *Post-BERT* takes 1 in 2019 and beyond and 0 before 2019. Only funds that used the machine download from 2015-2017 are included in this table; funds that did not use the machine download for their annual reports during this time period are excluded. The data used in this table spans the period 2015-2022 All regressions contain industry-year, stock, and hedge fund-year fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

	(1)	(2)
Dep Variable	Position Change	
Model:	BERT	LM Dictionary
Sentiment (MD&A) ×Post-BERT	0.032**	-0.417
	(0.012)	(0.281)
Sentiment (MD&A)	-0.013	1.150**
	(0.012)	(0.418)
Controls	Yes	Yes
Year×Industry FE	Yes	Yes
Stock FE	Yes	Yes
Hedge Fund FE	Yes	Yes
Obs	933,623	933,623
Adj R^2	0.167	0.0162

Table 9 Hedge Fund Trading Based on Textual: Sub-sectional analysis

This table examines the impact of the textual index of sub-sections in annual reports on hedge fund portfolio rebalancing. The sample in this table contains all position changes of machine download funds that have downloaded the companies' 10-K in that quarter. The *Index* variable in the six columns is listed as follows: *Sentiment* in the first column, *Uncertainty* in the second column, *Litigation* in the third column, *Modal Strong* in the fourth column, *Modal Weak* in the fifth column, and *Financial Constraint* in the sixth column. *Index (Risk)*s in Panel A are constructed based on the text in the Item 1A "Risk Factors" section. *Index (MD&A)*s are constructed based on the text in Item 7 "Management's Discussion and Analysis of Financial Condition and Results of Operations" section. We also control the text index of the full text of the annual report, *Index (Full text)*, so that the text index of the subsections captures the net impact of the text information of the subsections. The dictionaries for each category of words are from Loughran and McDonald (2011) and Loughran and McDonald (2015). All regressions contain industry-year, stock, and hedge fund fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

Panel A: Risk-Disclosure (Item-1A)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep variable:			Position Change			
Index:	Sentiment	Uncertainty	Litigation	Modal Strong	Modal Weak	Fin Constraint
Index (Risk)	-0.414	-0.096	1.303*	-3.266	1.557	0.958
	(0.525)	(0.082)	(0.619)	(3.442)	(0.982)	(1.158)
Index (Full text)	0.753***	-1.301**	-0.962	-5.488**	-1.650**	0.005
	(0.287)	(0.589)	(1.114)	(2.331)	(0.657)	(1.630)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	523,792	523,792	523,792	523,792	523,792	337,057
Adj R^2	0.128	0.128	0.128	0.128	0.128	0.158

Panel B: MD&A (Item-7)

	(1)	(2)	(3)	(4)	(5)	(6)
Dep variable:			Position Change			
Index:	Sentiment	Uncertainty	Litigation	Modal Strong	Modal Weak	Fin Constraint
Index (MD&A)	2.309***	-1.123*	-0.369	-4.799**	-1.530*	-0.201
	(0.331)	(0.515)	(0.386)	(1.950)	(0.809)	(1.181)
Index (Full text)	0.400	-1.633**	-0.315	-8.668**	-2.039***	-1.047
	(0.544)	(0.572)	(1.148)	(3.333)	(0.570)	(0.893)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	517,374	517,374	517,374	517,374	517,374	332,953
Adj R^2	0.129	0.128	0.128	0.128	0.128	0.159

Table 10 Cross-sectional variances vs. time series variances

This table examines the impact of time-series differences in text indexes on hedge fund position adjustments. The sample in this table contains all position changes of machine download funds that have downloaded the companies' 10-K in that quarter. *ΔIndex* is given by $\Delta Index_{i,t} = \frac{Index_{i,t} - Index_{i,t-1}}{Index_{i,t-1} \times 100}$, where *i* indicates stock and *t* indicates the year. The *Index* variable in the four columns is listed as follows: *Sentiment* in the first column, *Uncertainty* in the second column, *Modal Strong* in the third column, and *Modal Weak* in the fourth column. The dictionary for each category of words is from Loughran and McDonald (2011). All regressions contain industry-year, stock, and hedge fund-year fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

	(1)	(2)	(3)	(4)
Dep variable:	Position Change			
Index:	Sentiment	Uncertainty	Modal Strong	Modal Weak
ΔIndex	0.440	-0.057	1.019	-0.019
	(0.491)	(0.393)	(2.343)	(0.529)
Index	0.750**	-1.613**	-7.654**	-1.967**
	(0.353)	(0.612)	(3.206)	(0.625)
Controls	Yes	Yes	Yes	Yes
Year×Industry FE	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes
Obs	485,818	485,818	485,818	485,818
Adj R^2	0.124	0.124	0.124	0.124

Table 11 The Performance of Trades based on 10-K Text

This table reports the performance of value-weighted portfolios formed on text indices. Six portfolios are constructed as follows: 1) long stocks with *Sentiment* above the median and short stocks with *Sentiment* below the median; 2) long stocks with *Uncertainty* below the median and short stocks with *Uncertainty* above the median; 3) long stocks with *Litigation* above the median and short *Litigation* below the median; 4) long stocks with *Modal Strong* below the median and short stocks with *Modal Strong* above the median; 5) long stocks with *Modal Weak* below the median and short *Modal Weak* above the median; 6) long stocks with *Financial Constrain* below the median and short stocks with *Financial Constrain* above the median. Only stocks whose annual reports have been downloaded by machine in the disclosure quarter are included in our portfolio. Each portfolio is rebalanced on June 30 of each year. Within each portfolio, a zero-investment strategy is formed, with long positions with a total value of 1 and short positions with a total value of -1, weighted by market capitalization. Portfolio performance is measured by the mean monthly excess return, risk-adjusted returns using CAPM, alpha of the Fama-French three-factor model (FF-3), and alpha of the Fama-French-Carhart four-factor model (FFC-4). Panel A uses the time interval from June 2000 to December 2022, while Panel B uses the time interval from June 2011 to December 2022 for column (1) to column (5) and from June 2015 to December 2022 for column (6). Newey-West adjusted standard errors are reported in parentheses. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

Panel A: 2003-2022

	(1)	(2)	(3)	(4)	(5)	(6)
	Sentiment	Uncertainty	Litigation	Modal Strong	Modal Weak	Fin Constraint
Excess Return	0.604***	0.658***	0.225	0.681***	0.658***	0.559*
	(0.201)	(0.200)	(0.204)	(0.206)	(0.201)	(0.203)
CAPM Alpha	0.424***	0.441***	0.246	0.451***	0.441***	0.334*
	(0.158)	(0.167)	(0.207)	(0.169)	(0.167)	(0.170)
FF-3 Alpha	0.438***	0.451***	0.251	0.461***	0.450***	0.344**
	(0.153)	(0.159)	(0.206)	(0.165)	(0.160)	(0.166)
FFC-4 Alpha	0.414***	0.418***	0.242	0.426***	0.417***	0.310*
	(0.154)	(0.159)	(0.206)	(0.164)	(0.159)	(0.166)
# Months	258	258	258	258	258	258

Panel B: From Publication to 2022

	(1)	(2)	(3)	(4)	(5)	(6)
	Sentiment	Uncertainty	Litigation	Modal Strong	Modal Weak	Fin Constraint
Excess Return	0.958***	1.060***	0.459	1.089***	1.059***	1.120**
	(0.367)	(0.364)	(0.375)	(0.375)	(0.365)	(0.556)
CAPM Alpha	0.369*	0.349*	0.540	0.340*	0.347*	0.384
	(0.208)	(0.200)	(0.352)	(0.169)	(0.201)	(0.304)
FF-3 Alpha	0.268*	0.263**	0.543	0.253**	0.261**	0.271
	(0.159)	(0.118)	(0.354)	(0.121)	(0.118)	(0.165)
FFC-4 Alpha	0.269	0.250**	0.504	0.242**	0.248**	0.270
	(0.175)	(0.123)	(0.371)	(0.124)	(0.122)	(0.164)
# Months	138	138	138	138	138	78

Appendix

Table A1 Variable Definitions

This table reports the definitions of the variables. The dictionaries for each category of words are from Loughran and McDonald (2011) and Loughran and McDonald (2015).

Variable	Definition
Position Change	Change in holdings of stock in a fund's position, when the position increases, is calculated as $Position\ Change_{i,j,t} = \frac{Stock\ Postion_{i,j,t} - Stock\ Postion_{i,j,t-1}}{Stock\ Postion_{i,j,t}}$, and when the position decreases, is calculated as $Position\ Change_{i,j,t} = \frac{Stock\ Postion_{i,j,t} - Stock\ Postion_{i,j,t-1}}{Stock\ Postion_{i,j,t-1}}$.
Sentiment	The difference between the number of positive words minus the number of negative words in the text of the annual report divided by the total number of words.
Uncertainty	The number of uncertain words in the annual report text divided by the total number of words.
Litigation	The number of litigation words in the annual report text divided by the total number of words.
Modal Strong	The number of modal strong words in the annual report text divided by the total number of words.
Modal Weak	The number of modal weak words in the annual report text divided by the total number of words.
Fin Constraint	The number of financial constraint words in the annual report text divided by the total number of words.
Negative_Harvard	The number of negative words defined by Harvard IV-4 dictionary in the annual report text divided by the total number of words.
Return(-1)	The stock's return for the previous quarter, calculated using the closing price at the end of the quarter and the closing price at the end of the previous quarter.
Vol(-1)	Stock's monthly volatility over the last quarter.
Size	Natural logarithm of the company's total assets.
B/M	The book value of the company divided by the market value of the company
ROE	The company's return on equity, calculated by dividing net income by shareholders' equity.
Investment	The increase in the company's assets in the previous fiscal year divided by the total assets at the end of the previous two fiscal years
Total words	Total number of words in the text of the annual report
Sue	Standard unexpected earnings of the stock, calculated by $Sue_t = \frac{EPS_t - E(Forcasted\ EPS_t)}{SD(Forcasted\ EPS_t)}$.
Excess Return	The mean monthly return over the risk-free rate.
CAPM Alpha	Risk-adjusted returns using the CAPM.
FF-3 Alpha	Risk-adjusted returns using the Fama-French three-factor model.
Carhart-4 Alpha	Risk-adjusted returns using the Carhart four-factor model.

Table A2 Decomposition of the Sentiment Index

This table reports regressions of quarterly Position Change of positions on different text indices. The unit of observation is a hedge fund-quarter-stock holding. The dependent variable for all columns is *Position Change*. The sample contains all positions of machine download funds that have downloaded the companies' 10-K in that quarter. *Positive* is the frequency of positive sentiment words in the annual report, and *Negative* is the frequency of negative sentiment words in the annual report. The independent variables used in column (1) and column (2) are constructed based on the entire text of the annual report. The independent variables used in column (3) and column (4) are constructed based on the text in Item 1A "Risk Factors" section. The independent variables used in column (5) and column (6) are constructed based on the text in Item 7 "Management's Discussion and Analysis of Financial Condition and Results of Operations" section. The dictionary for each category of words is from Loughran and McDonald (2011). All regressions contain industry-year, stock, and hedge fund fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

	(1)	(2)	(3)	(4)	(5)	(6)
Section:	Full Text		Risk-Disclosure		MD&A	
Positive	1.439		1.624		3.457**	
	(1.190)		(1.097)		(1.062)	
Negative		-1.189***		0.369		-2.042***
		(0.300)		(0.306)		(0.384)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	540,213	540,213	523,792	523,792	517,374	517,374
Adj R^2	0.128	0.128	0.128	0.128	0.129	0.129

Table A3 Consider Earnings Surprise

This table reports regressions of quarterly Position Change of positions on different text indices. The unit of observation is a hedge fund-quarter-stock holding. The dependent variable for all columns is *Position Change*. The sample contains all positions of machine download funds that have downloaded the companies' 10-K in that quarter. *Sue* is the standard unexpected earnings given by $Sue_t = \frac{EPS_t - E(Forcasted\ EPS_t)}{SD(Forcasted\ EPS_t)}$, where EPS_t is the earnings per share reported in an annual report, $E(Forcasted\ EPS_t)$ is the forecasted or anticipated earnings per share for a company during the same fiscal year and $SD(Forcasted\ EPS_t)$ is the standard deviation of estimated earnings for the fiscal year. The dictionary for each category of words is from Loughran and McDonald (2011). All regressions contain industry-year, stock, and hedge fund fixed effects. Standard errors are two-way clustered at the stock and year levels and reported in parentheses. Statistical significance is represented by * p < 0.10, ** p < 0.05, and *** p < 0.01.

	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment	1.005**					
	(0.369)					
Uncertainty		-1.865**				
		(0.473)				
Litigation			0.204			
			(0.643)			
Modal Strong				-6.982**		
				(3.135)		
Modal Weak					-2.278**	
					(0.472)	
Fin Constraint						-0.388
						(1.865)
Sue	0.004	0.004	0.004	0.004	0.004	0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year×Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Stock FE	Yes	Yes	Yes	Yes	Yes	Yes
Hedge Fund FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	481,526	481,526	481,526	481,526	481,526	233,137
Adj R^2	0.130	0.130	0.130	0.128	0.128	0.180