Summer 7-15-2021

Uncovering Mutual Fund Private Information with Machine Learning

Liang Zhang

doi: https://doi.org/10.57709/24069435

Uncovering Mutual Fund Private Information with Machine Learning

BY

Liang Zhang

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree

Of

Doctor of Philosophy

In the Robinson College of Business

Of

Georgia State University

GEORGIA STATE UNIVERSITY

ROBINSON COLLEGE OF BUSINESS

2021

ACCEPTANCE

This dissertation was prepared under the direction of the Liang "Alan" Zhang Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business Administration in the J. Mack Robinson College of Business of Georgia State University.

Richard Phillips, Dean

DISSERTATION COMMITTEE

Dr. Vikas Agarwal

Dr. Baozhong Yang

Dr. Zhen Shi

Dr. Sean Cao (external – School of Accountancy)

Dr. Wei Jiang (external – Columbia Business School)

ACKNOWLEDGEMENT

This dissertation and Ph.D. degree would not have been possible without the help, support, and guidance of many people. I thank my dissertation committee of Vikas Agarwal, Sean Cao, Wei Jiang, Zhen Shi, and Baozhong Yang for making me a better academic researcher. I would like to give special thanks to my advisor Vikas Agarwal for his continued support and encouragement. I am deeply grateful to Baozhong Yang for being a mentor and friend since my first day in the Ph.D. program. I appreciate all finance faculty for their attention and help throughout the years. I thank my fellow Ph.D. students for their friendship during this journey.

I am forever grateful to my family for their care, understanding, and encouragement. I dedicate this dissertation to my parents, Zhaohui Zhang and Jie Zhao, for their everlasting love.

ABSTRACT

Uncovering Mutual Fund Private Information with Machine Learning

BY

Liang Zhang

July 28, 2021

Committee Chair: Dr. Vikas Agarwal

Major Academic Unit: Department of Finance

This paper implements natural language processing (NLP) models and neural networks to predict mutual fund performance using the textual information disclosed in mutual fund shareholder letters. Informed funds identified by the prediction model deliver superior abnormal returns and are more likely to receive an upgrade in Morningstar ratings. Informed funds also attract greater flows in three days and up to 24 months after the disclosure of shareholder letters, especially when their disclosure has greater investor attention, suggesting that investors recognize the information from the qualitative disclosure. The machine learning model shows that informed funds tend to discuss sector specializations, portfolio risk taking, big picture of the financial market, and mixed strategies across assets. Collectively, this study shows that mutual fund disclosure contains rich, value-relevant textual information that can be analyzed by state-of-the-art machine learning models and help investors identify informed funds.

Uncovering Mutual Fund Private Information with Machine Learning

Alan L. Zhang*

July 2021

ABSTRACT

This paper implements natural language processing (NLP) models and neural networks to predict mutual fund performance using the textual information disclosed in mutual fund shareholder letters. Informed funds identified by the prediction model deliver superior abnormal returns and are more likely to receive an upgrade in Morningstar ratings. Informed funds also attract greater flows in three days and up to 24 months after the disclosure of shareholder letters, especially when their disclosure has greater investor attention, suggesting that investors recognize the information from the qualitative disclosure. The machine learning model shows that informed funds tend to discuss sector specializations, portfolio risk taking, big picture of the financial market, and mixed strategies across assets. Collectively, this study shows that mutual fund disclosure contains rich, value-relevant textual information that can be analyzed by state-of-the-art machine learning models and help investors identify informed funds.

JEL Classification: C45, G11, G14, G23

Key Words: Machine Learning, Mutual Fund Performance, Disclosure, Textual Analysis, Shareholder Letters, Fund Flows

*J. Mack Robinson College of Business, Georgia State University. Email: lzhang27@gsu.edu. I am grateful to my dissertation committee members: Vikas Agarwal, Sean Cao, Wei Jiang, Zhen Shi, and Baozhong Yang. I have benefited from the discussions with Merlin Bartel, Jillian Grennan, Bin Ke, Ville Rantala (discussant), David Reeb, Elvira Sojli (discussant), Stefan Voigt (discussant), Junbo Wang, and Xinde Zhang (discussant), and comments and suggestions from participants in seminars and conferences at Machine Learning and Business Conference at University of Miami, Conference on Financial Innovation at Stevens Institute of Technology, 2021 China International Risk Forum, 11th Financial Markets and Corporate Governance Conference PhD Symposium, 2nd Lixin Conference on New Frontiers in Finance, Florida International University and Georgia State University. The usual disclaimer applies.

1. Introduction

For investment companies (e.g., mutual funds), shareholder reports are more than a legal requirement to disclose information such as portfolio holdings, fund performance, accounting statements, and voting policies. It offers an effective channel to communicate with shareholders as well as potential investors on various topics, including dissection of wins and losses, comments on sector and fund performance, emphasis of the investment philosophy, and insights on the state of the economy and market. In his shareholder report on August 24, 2016, the president of Impax Asset Management, Joseph Keefe's concern on politics expanded such an unconfined list of topics as he wrote at the beginning of his report: "To top things off, we have Brexit ... and a wave of ... anti-immigration, anti-globalization sentiments suddenly sweeping western countries, further clouding the economic horizon." However, he remained confident in his funds and added: "We strongly believe that over time our shareholders benefit from investing in companies that meet higher environmental, social and governance (ESG) standards. ... We invest for the long term, so it is slow work, but it is vital work. It may not make the headlines, but it is making a difference."

While almost all quantitative information from shareholder reports originates from portfolio holdings, which have been long studied by investors, analysts, and researchers, the qualitative data from the rich textual discussions (i.e., shareholder letters[1]) remain underexplored.[2] Do managers share their insights and privileged information with the public in shareholder letters? Can investors learn valuable information, such as managers' investment skill, from the disclosure in shareholder letters? Which funds benefit from such disclosures? This paper tackles these questions.

There are several empirical challenges in extracting value-relevant information from mutual fund shareholder letters. The first hurdle is to extract intrinsic syntactic and semantic

[1] In mutual fund shareholder reports, managers generally include shareholder letters to provide qualitative discussions on various topics such as fund performance, sector performance, market overview, risk taking, fiscal policy, politics, and global issues.

[2] For example, see "Qualitative Guidelines for Mutual Fund Selection," John Deysher and Michael Walters, June 2014, American Association of Individual Investors.

features from the unstructured text. Traditional bag-of-words approach in textual analysis relies on the meaning of individual vocabulary words and thus omits higher-order interactive features among words and sentences, which can contain important qualitative information. For example, the word "board" would have the same context-free representation in "welcome on board" and "board of directors." We overcome this challenge by implementing one of the most cutting-edge developments in natural language processing (NLP) – Bidirectional Encoder Representations from Transformers (BERT) created and developed by Devlin, Chang, Lee, and Toutanova (2019). Unlike traditional language representation models that read a context either from left to right or right to left, BERT jointly conditions on both left and right contexts. Because BERT pre-trains deep bidirectional representations from a large quantity of unlabeled texts, it can capture the higher-order semantic and syntactic structure among words and sentences without losing information from a context.

The second obstacle arises from decoding relevant features from shareholder letters, i.e., determining what features are likely to be associated with the private information of fund managers. We resolve this issue by building a recurrent neural network model to learn the relation between linguistic features (extracted from fund managers' shareholder letters via BERT) and subsequent fund performance (computed as the alpha from the Fama and French (1993) and Carhart (1997) four-factor model). To train and validate our model, we split the sample of shareholder letters retrieved from the Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system into a training set and a test set. We train the model using shareholder letters from 2006 to 2014 (i.e., the training set) to fit it in financial contexts and then use the trained model to predict future fund performance based on shareholder letters from 2015 to 2018 (i.e., the test set). We conduct all empirical analyses in the test period to avoid using any information from the training process.

We define a text-based fund informativeness measure, *Textual Fund Information*, based on the predictions by the neural network model, and classify funds into informed and unin-

formed funds. We show that informed funds identified from the information in shareholder letters outperform uninformed funds by 0.9% every year. Even after controlling for various fixed effects, fund characteristics, and skill measures that prior literature has shown to be correlated with future fund performance,[3] the economic and statistical significance of the outperformance of informed funds is virtually unchanged. These findings suggest that *Textual Fund Information* is a time-varying measure and captures information that is largely independent of observable fund characteristics and time-invariant unobservables at manager, fund, and company levels. To account for the possibilities that managers may hide bad information or have different writing styles, we control for the textual measures including sentiment, readability and complexity, and the results remain intact.

Recent studies (Ben-David, Li, Rossi, and Song, 2019; Cheng, Lu, and Zhang, 2021; Evans and Sun, 2021) document that mutual fund investors allocate their capital flows based on Morningstar fund ratings. 7% of mutual funds in our sample explicitly discussed Morningstar ratings in 2018. Given the impact of Morningstar ratings on funds and especially on investors, we investigate whether *Textual Fund Information* is positively associated with Morningstar ratings and the likelihood to receive an upgrade in ratings subsequently. We find that not only do informed funds have higher future Morningstar ratings (i.e., level of ratings), but they are also more likely to be rewarded higher ratings (i.e., change of ratings). These findings further underscore the importance of *Textual Fund Information* for investors on account of its predictive power on Morningstar ratings.

Given that the textual information in mutual fund disclosure predicts future fund performance, a natural question is whether investors can identify informed funds from their disclosure and invest in such funds. Since shareholder letters are generally filed semi-annually, we relate a fund's *Textual Fund Information* to its capital flows over the six months after investors receive shareholder letters. We show that informed funds receive 1.8% more capital

[3]The fixed effects include manager, fund, and company fixed effects. The skill measures include return gap (Kacperczyk, Sialm, and Zheng, 2008), DGTW benchmark-adjusted return (Daniel, Grinblatt, Titman, and Wermers, 1997), and portfolio holdings return.

inflows than uninformed funds over the six-month period. To the extent that fund skill will be revealed in the long run, we further examine flows in the one year and two years after investors receive shareholder letters to account for the possibility that uninformed funds manipulate the contents in shareholder letters to reduce the short-term redemption risk. We find that informed funds attract 3% and 4.6% more flows than uninformed funds over one year and two years, respectively. To reduce the concern that investors chase the most recent performance, we control for contemporaneous performance when examining changes in flows over longer horizons. The semi-annual disclosure frequency raises the possibility that concurrent new information from other sources, rather than the information in shareholder letters, influences investors' decisions. We thus employ daily flow data from Trimtabs database for a subset of funds and confirm that investors put more capital into informed funds within 3-day and 5-day windows once they receive shareholder letters. For example, informed funds attract about 25 bps more flows than uninformed funds within five days, equivalent to a dollar value of $4.48 million for an average fund.

Since the target recipients of shareholder letters are investors, we consider how investors' behavior affects fund managers' expectations of disclosing information. Although it is implausible to observe how each investor reads the letters and allocates her investment, we use the number of visits to a shareholder letter on the SEC EDGAR website as a proxy for the readership and investor attention on the letter. We show that informed funds attract more flows in both the short term and long term when there exists a readership on their shareholder letters. On the other hand, when such investor attention is nonexistent, there is no difference in flows between informed and uninformed funds. This finding establishes a direct link between *Textual Fund Information* and the subsequent investors' actions.

We next endeavor to understand what contents are discussed by informed funds and which funds are likely to include quality information in shareholder letters. Understanding the first question helps us uncover the informational contents that contribute to the predictive power of *Textual Fund Information*. We employ topic modeling to classify shareholder letters

into four clusters via Latent Dirichlet Allocation (LDA) developed by Blei, Ng, and Jordan (2003). Based on the top five keywords for each cluster, we find informed funds tend to discuss sector specialization, portfolio risk taking, the big picture of the financial market, and strategies in mixed assets. Next, we show that our neural network model achieves greater accuracy in identifying informed funds among those with a riskier portfolio. Informed funds with a riskier portfolio generate superior performance, attract more capital flows, and have a greater likelihood of receiving upgraded Morningstar ratings. In other words, these funds tend to provide more quality information to investors because their actual skill is not easily distinguishable from luck. Therefore, our model adds unique value on assessing hard-to-evaluate opaque funds.

Given the topic-free nature of shareholder letters, some funds choose not to provide any subjective discussion. Instead, they report quantitative information readily available from other data sources such as performance in the last period or historical performance since a fund's inception. We do not include these shareholder reports in our sample because of the lack of any qualitative information. However, it would be instructive to understand which kind of funds are likely to provide active discussion. A determinant model built on the decision of including a shareholder letter or not indicates that fund managers who take higher risk, exhibit greater portfolio turnover, charge higher marketing expenses (12b-1 fees), and seek for external labor market opportunities are more likely to utilize shareholder letters as an effective tool to communicate with current and potential investors. These findings suggest that funds that stand to benefit more from the additional qualitative disclosure in shareholder letters (e.g., more transparency and publicity) choose to do so.

The contribution of this study is three-fold. First, this paper adds to the literature on mutual fund skill and performance. Cremers and Petajisto (2009) and Amihud and Goyenko (2013) identify skilled mutual funds based on measures constructed from historical fund returns. Other studies on fund performance use fund holdings data. Brands, Brown, and Gallagher (2005), Kacperczyk, Sialm, and Zheng (2005, 2008), and Cremers, Ferreira, Matos,

and Starks (2016) show that active management and stock selectivity enhance fund performance. Daniel, Grinblatt, Titman, and Wermers (1997) show that stocks picked by mutual funds outperform a characteristic-based benchmark. Kacperczyk and Seru (2007) find that funds that use less public information for selecting stocks exhibit better performance. Ferson and Mo (2016) develop holdings-based performance measures to identify funds with security selectivity. Our study departs from this literature by developing a new methodology to measure fund private information and, to our knowledge, is the first study to create such a measure based entirely on textual contents of mutual fund disclosures.

We also contribute to the burgeoning literature on the textual analysis by introducing advanced machine learning techniques. The application of textual analytic techniques has benefited research in finance, accounting, and economics (see survey articles, e.g., Loughran and McDonald, 2016; Gentzkow, Kelly, and Taddy, 2019. The text representation from bag-of-words models widely used in the literature (e.g., Hillert, Niessen-Ruenzi, and Ruenzi, 2016) only captures context-free features of individual words and fails to account for the sentence, the sequence of words, and relations among words. Other studies (e.g., Ke, Manela, and Moreira, 2019; Garcia, Hu, and Rohrer, 2020) use machine learning algorithms to classify the sentiment of news articles and conference call transcripts. We employ an advanced NLP model to capture higher-order semantic and syntactic features from mutual fund managers' qualitative discussion. In addition, we apply a deep neural network model to translate these features into fund information and interpret informational contents on which informed funds are likely to discuss in their shareholder letters. Our study suggests that machine learning techniques can assist in studying unstructured data that contain rich information and providing insights into finance research.

Finally, our study contributes to the literature on disclosure. Theoretical literature documents that managers' equilibrium disclosure strategies depend on the nature of competition and their information environment (e.g., Diamond and Verrecchia, 1991; Admati and Pfleiderer, 2000). Cao, Du, Yang, and Zhang (2021) find that it requires stock picking skill

and intensive research on firm fundamentals to reap benefit from copycatting competitors' disclosed holdings. Since the shareholder letters contain less straightforward information than disclosed holdings and is more difficult and requires more efforts to copycat, our setting attenuates the possibility that managers may hide information due to the concern about their competitors. While the literature on corporate disclosure often finds that disclosing information is associated with proprietary costs,[4] our findings shed lights on the potential benefits of disclosure in the mutual fund industry. We provide empirical evidence that fund managers can send their private signal by voluntarily disclosing qualitative information to investors and find that informed managers, especially those whose skill is difficult to capture through their historical performance, can benefit from disclosures. Moreover, we complement the existing literature to show that disclosures are beneficial when target recipients of disclosures are likely to materialize the information that they receive.

2. Applying NLP and Neural Networks on Shareholder Letters

We implement two major steps to convert shareholder letters into a textual skill measure, *Textual Fund Information*, ready to use for our empirical analysis. The first step is to extract features from shareholder letters, and the second one is to build a neural network model that uses extracted features as input and generates a prediction of informativeness for each fund.

The informal format of the shareholder letters conveys rich information to investors but creates a hurdle for researchers because of the unstructured nature. The traditional bag-of-words approach in the textual analysis may omit useful information. For illustration purposes, here is a hypothetical discussion: "The fund performance was not bad." In this example, the manager should have a positive or neutral tone on fund performance, but the bag-of-words approach would consider the tone to be negative.

There are two dimensional features from textual information. Lexicalized features and

[4]See, for example, Ali, Klasa, and Yeung (2014), Bernard (2016), Cao, Ma, Tucker, and Wan (2018), and Li, Lin, and Zhang (2018).

higher-order interaction features among words. The former captures each word's meaning and property while the latter captures the sequence of words and relations between words (Chen and Manning, 2014). However, as Loughran and McDonald (2016) point out, the bag-of-words approach may not work well in the above example because of its limitation on analyzing the semantic and syntactic structure among words.

We overcome the limitation by applying one of the most path-breaking developments in NLP – BERT. It is the first bidirectional and unsupervised language representation and is a neural network-based technique for NLP. Traditional NLP models, such as *word2vec* and *GloVe*, are context-free and produce a single word embedding representation for each word in the vocabulary. Moreover, these models read a text sequence either from left to right or right to left.

However, not only does BERT read the entire text bidirectionally, but it also catches all surroundings of a word to provide a context-dependent representation for the word. BERT is trained on entire Wikipedia with 2,500 million words and the Books Corpus with 800 million words. We apply the pre-trained base BERT model on shareholder letters and extract the aggregate representation of a sentence in the final hidden state, denoted as a vector $C \in \mathbb{R}^{\mathrm{H}}$ where H is the hidden size and equals to 768 in the base BERT model. High-quality features C later serve as the input for a neural network model designed to select informed mutual funds (see Appendix B for details of BERT).

One limitation of BERT is the applicability on inputs longer than several hundred words (Pappagari, Zelasko, Villalba, Carmiel, Dehak, 2019). The length of shareholder letters spans from less than a hundred words to over 2,000 words, resulting in applying BERT on the entire document to be computationally inapplicable. To accommodate the applicability of BERT, for each shareholder letter, we first use BERT on every sentence and then create an $N \times C$ matrix as the aggregate representation of a shareholder letter, where N is the number of sentences in a letter.

In the second major step, we build a recurrent neural network model in Keras, an open-

source library that provides a Python interface for artificial neural networks. Our model contains four hidden layers with an additional Long Short-Term Memory (LSTM) layer.

We split all shareholder letters from 2006 to 2018 into a training set (from 2006 to 2014) and a test set (from 2015 to 2018). For the training set, we compute the future fund Alpha (see Section 3.2 for variable construction) after each shareholder letter is filed. Next, each year, we sort letters into tercile based on *Alpha* and label them with 2 (top tercile), 1, and 0 (bottom tercile). In other words, letters with the label of 2 represent outperforming funds in a year, while letters with 0 consist of underperforming funds in the same year. We train our model with features generated from BERT and labels created from *Alpha* using all shareholder letters in the training set.

After the model is trained, we apply the model on shareholder letters in the test set and generate a predicted label for each shareholder letter.[5] For example, a fund's shareholder letter with a predicted label of 2 (0) indicates that the fund would rank in the top (bottom) tercile in terms of future *Alpha*.

Shareholder letters can have two components, backward looking information and forward looking . Our research design helps extract the forward looking component, which has unique value to researchers and investors. In the training set, the neural network model only captures forward looking component that is predictive of future performance and effectively omits the backward looking component. When the predictability is carried over to the test set, *Textual Fund Information* is, therefore, constructed only by the forward looking component.

[5]We first train shareholder letters from 2006 to 2014 and predict labels for letters in 2015. We then expand the training set with letters in 2015 and predict labels for letters in 2016 with the updated training set. We repeat this rolling procedure until the end of the sample.

3. Data, Variable Construction, and Sample Overview

3.1. Mutual fund shareholder reports

We scrape mutual funds' shareholder reports from N-CSR (certified annual shareholder reports for management investment companies) and N-CSRS (certified semi-annual shareholder reports for management investment companies) filings on the SEC EDGAR website from 2006 to 2018. A registered investment company (e.g., a mutual fund company) must electronically file Form N-CSR (We use N-CSR to represent both N-CSR and N-CSRS filings thereafter) to the SEC within ten days of sending corresponding reports to shareholders.

In many cases, managers include shareholder letters covering managers' qualitative discussions on various topics such as fund performance, sector performance, market overview, risk taking, fiscal policy, politics, and global issues. Unlike other parts in Form N-CSR, these letters do not follow any template and therefore serve as effective communications between funds and their investors. However, there does not exist a clear-cut section for shareholder letters. We, therefore, write a computer program in Python to locate the beginning and the end of a letter by common phrases and complement with manual work if the computer program fails to extract the letter.[6]

As discussed in Section 2, shareholder letters from 2015 to 2018 serve as the test set for the neural network predictions and are used for our empirical analysis. Based on the labels of shareholder letters generated by our neural network model, we consider that funds whose shareholder letters receive a label of 2 as predicted informed funds and those with a label of 0 as predicted uninformed funds. Funds with labels of 2 or 0 serve as the primary sample in our empirical analysis.[7] Specifically, the key variable of interest in this study, *Textual Fund*

[6]We first manually read 100 filings and summarize a list of common phrases such as "Dear Shareholders" and "Sincerely;" then we apply the computer program on all filings. For filings that the program fails to locate a letter, we manually check whether they do not have a letter or they use other phrases for the beginning and the end of the letter; for the latter, we expand our list of common phrases and reapply the program. Finally, we manual check the precision of letter extractions.

[7]The difference between predicted informed and uninformed funds provide more direct and sharp comparison on the information contents of the shareholder letters. Nevertheless, we include funds predicted to perform neutrally (with a label of 1) as robustness checks and results are qualitatively similar.

Information is an indicator variable equal to one if a fund is predicted to be informed and zero if it is predicted to be uninformed based on shareholder letters.

To obtain data on the readership of shareholder letters, we use SEC EDGAR's associated Log File Dataset which tracks the traffic of requests and downloads. Specifically, it comprises all records of the requests of SEC filings on EDGAR system from January 2003 to June 2017. Each observation in the original dataset contains information on the visitor's Internet Protocol (IP) address, timestamp, and the identifier of the filing that the visitor downloads. Because shareholder reports are a part of Form N-CSR, we use the number of downloads on a Form N-CSR to proxy for the readership on the embedded shareholder letters. We keep all non-duplicate requests, including those made by robots and algorithms since recent studies (Cao, Du, Yang, and Zhang, 2021; Cao, Jiang, Yang, and Zhang, 2021) show that the information acquisition by automated downloaders exerts significant effects on disclosing firms and funds.

3.2. Mutual fund data

We obtain fund return data and fund characteristics such as expense ratio, turnover ratio, total net assets (TNA), and fund age from the Center for Research in Security Prices (CRSP) Survivorship Bias Free Mutual Fund database and fund portfolio holdings from the Thomson Reuters Mutual Fund Holdings (s12) database. We use the MFLINKS tables provided by Wharton Research Data Services (WRDS) to merge the CRSP Mutual Fund database and the Thomson Reuters s12 database. In addition, for more in-depth analyses on the timing of capital flows, we use daily flow data from Trimtabs database.

To merge the N-CSR shareholder reports and the mutual fund databases, we construct a link between Series ID (fund identifier in N-CSR) and the WFICN (Wharton Financial Institution Code Number; the identifier for fund portfolios in MFLINKS). Beginning on February 6, 2006, all open-ended mutual funds are required by SEC to report series (fund portfolio) and class (share class) identification information in their N-CSR filings. For each Series ID,

mutual fund companies also report the underlying share class information, including Class ID, Class Name, and Class Ticker. We use the Class Ticker to match with the ticker symbol in the CRSP Mutual Fund database. When a share class is matched by a ticker symbol, we consider the associated Series ID and WFICN as matched. Since Series ID and WFICN are both fund portfolio level identifiers, we drop the cases in which one Series ID is matched to multiple WFICNs. At the portfolio level, we are able to match N-CSR filings with CRSP data for 2,910 domestic equity mutual funds.[8]

Although mutual funds start to file N-CSR in 2003, the series and class identification information are not mandatory until 2006. Therefore, we use filings from January 2006 to December 2018. Over the 13-year span, our initial sample consists of 17,717 N-CSR filings with shareholder letters by domestic equity funds.

Since we are interested in managers' skill in making investment decisions, we drop ETFs, annuities, and index funds and focus on actively managed funds. In addition, we follow the conventional selection criteria in Kacperczyk, Sialm, and Zheng (2008) to identify domestic equity funds. We aggregate all share classes at the fund level. *TNA* is aggregate total net assets ($mm) across all share classes one month before a filing date. *Age* is the number of years since the fund's oldest share class is launched. We use the natural logarithms of *TNA* and *Age* in our empirical analyses. Return-based variables, turnover ratio (*Turnover*), expense ratio (*Expense*), 12b-1 fees (*12b1Fee*), and management fees (*MgmtFee*) are the TNA-weighted average across all fund share classes and scaled to percentage points. *Tenure* is the number of years since a portfolio manager is hired; if there are multiple managers for a fund, the longest tenure is used.

[Insert Table 1 Here]

In general, funds report shareholder reports semi-annually. We therefore use a 180-day (or 6-month) window after the filing of a shareholder report to measure fund performance

[8]Note that N-CSR is filed at the company level and a company may use the same filing for several funds within the fund family. In untabulated results, we find that the returns of funds under the same CIK are highly correlated. We also find qualitatively similar results in untabulated analyses when we exclude filings that contain reports for multiple funds in the same company.

(*Alpha*), computed as the intercept from the regression of daily excess returns on the Fama-French-Carhart four factors, annualized by multiplying with 253. We follow the conventional approach in Sirri and Tufano (1998) to construct flow measures in various windows, including 3-day, 5-day, 6-month, 12-month, and 24-month. *MSRating* is the fund rating assessed by Morningstar[9] and retrieved in six-month post filing. To control for past risk taking and performance, we use the daily returns in the 180 days prior to the filing of a shareholder report to construct the annualized *PastRisk*, defined as the standard deviation of daily returns multiplied by the squared root of 253, and *PastAlpha* measures. We control *PastFlow* by calculating flows over the 6-month prior to the filing. To control for concurrent information from portfolio holdings disclosed in the same Form N-CSR, we create two measures, *HoldingReturn* and *DGTWReturn*, defined as the raw return and DGTW benchmark-adjusted return on a fund's portfolio holdings in the six month ($\sim$ 180 days) post filing. All potentially unbounded variables are winsorized at the 1% extremes.

3.3. Textual measures of shareholder letters

The traditional textual analysis uses a bag-of-words approach to measure managers' sentiment in the letter. *LM_Negative* (*LM_Positive)* is the number of Loughran-McDonald (Loughran and McDonald, 2011) finance-related negative (positive) words in a shareholder letter divided by the total number of words in the shareholder letter, expressed in percentage points. To account for the readability and complexity of the writing, we use *DocumentSize*, defined as the natural logarithm of the number of words in a shareholder letter to proxy for readability (Loughran and McDonald, 2014), and *TextDiversity*, defined as the ratio of the number of distinct words in a shareholder letter to the square root of the total number of words to capture complexity (Carroll, 1964). We include these variables as textual control variables.

[9]For the complete methodology to calculate fund ratings by Morningstar, see: https://www.morningstar.com/content/dam/marketing/shared/research/methodology/771945_Morningstar_Rating_for_Funds_Methodology.pdf

4. Can *Textual Fund Information* Select Informed Funds?

4.1. Fund performance

One of the critical tasks is to confirm the joint hypothesis that mutual funds disclose private information in shareholder letters so that our neural network model can be successful in selecting informed funds. Therefore, we examine the future fund performance to validate the predictions from our model.

In this study, we focus on domestic equity funds because they are the predominant class of mutual funds, and their performance has well-defined benchmarks, such as the Fama-French-Carhart four-factor model. We consider the following regression at the filing level, indexed by fund(i)-filing(j)-date(t), with year, fund, fund family, and manager fixed effects:

$$Alpha_{i,j,t} = \beta Textual\ Fund\ Information_{\ i,j,t} + \gamma FundChar_{i,j,t} + \alpha_i + \alpha_{family} + \alpha_{manager} + \alpha_{year} + \epsilon_{i,j,t} \quad (1)$$

Under the hypothesis that the neural network model is competent in selecting informed funds, we expect coefficient β to be significantly positive. Results in Table 2 confirm such a conjecture. Informed funds identified by the prediction model beat uninformed funds across all specifications. The first four columns show that the annual abnormal return of informed funds is 80 to 88 bps higher. A comparison between the specifications with and without fund fixed effects is revealing. Adding fund fixed effects virtually does not affect β, suggesting that the informativeness of a fund is time varying and depends on the manager's private information for the next period. We further add manager fixed effects and find qualitatively similar results, confirming that *Textual Fund Information* captures managers' information set in the subsequent period and is not driven by unobserved, time-invariant manager and

company characteristics.[10]

[Insert Table 2 Here]

Because the training and predicting process only uses textual contents from shareholder letters, *Textual Fund Information* should be independent of other characteristics or skill measures considered by the existing literature to be correlated with future fund performance. Indeed, after controlling for fund characteristics such as past performance, size, and age, and skill measures[11], the economic and statistical significance of the outperformance of informed funds is virtually unchanged. Because of its time-varying nature, *Textual Fund Information* is distinct from skill measures documented in the extant literature. Managers also disclose portfolio holdings, which may contain information about future performance, along with shareholder letters in Form N-CSR, we further control for contemporaneous returns on portfolio holdings, *HoldingReturn* and *DGTWReturn*, to capture information orthogonal to portfolio holdings, and obtain similar coefficients on *Textual Fund Information*, suggesting that *Textual Fund Information* provides incremental information to portfolio holdings, which are presented in the same disclosure.

The effects are almost invariant with or without the textual control variables such as *LM_Negative*, indicating that traditional bag-of-words measures in the textual analysis have little confounding effect. These textual variables related to sentiment and tone also alleviate the concern that managers may hide bad information. Managers may also have different writing styles in the shareholder letters or use complicated language after poor performance, we thus use *DocumentSize* and *TextDiversity* to proxy for readability and complexity of writing, respectively, to rule out the possibility that writing styles affect the value-relevant information in the disclosure. The results are unaltered after accounting for the complexity

[10]In untabulated analysis, we control for manager switches and find that managerial turnover does not affect the predictive power of *Textual Fund Information*, suggesting that the personal writing style is unlikely to be correlated with *Textual Fund Information*.

[11]In untabulated analysis, we control for ex ante skill measures, including return gap (Kacperczyk, Sialm, and Zheng, 2008), DGTW benchmark-adjusted return (Daniel, Grinblatt, Titman, and Wermers, 1997), R-squared from the Fama-French-Carhart four-factor model (Amihud and Goyenko, 2013), and past portfolio holdings return

of language. Interestingly, managers who write longer letters with more concise language tend to perform better.

4.2. Morningstar fund rating

Recent studies (Ben-David, Li, Rossi, and Song, 2019; Cheng, Lu, and Zhang, 2021; Evans and Sun, 2021) document that investors use Morningstar fund ratings to allocate their capital flows. It is, therefore, rational for funds to improve their ratings. For instance, 7% of shareholder letters in 2018 explicitly mentioned Morningstar ratings or benchmark returns created by Morningstar.

We consider Morningstar rating as another proxy for skill and expect informed funds predicted by the neural network model to have higher future Morningstar ratings and be more likely to experience an upgrade in Morningstar ratings. Specifically, we consider the following regression at the filing level, indexed by fund(i)-filing(j)-date(t), with year and fund (or fund family) fixed effects:

$$MSRating_{i,j,t} = \beta TextualFundInformation_{i,j,t} + \gamma FundChar_{i,j,t} + \alpha_i(\alpha_{family}) + \alpha_{year} + \epsilon_{i,j,t} \tag{2}$$

Note that the coefficient β captures the relation between *Textual Fund Information* and the level of *MSRating*. If we add *PastMSRating* as a control variable, then the coefficient β captures the relation between *Textual Fund Information* and the change in *MSRating*.

[Insert Table 3 Here]

Table 3 reports the results for both cases. For example, column (1) shows that informed funds are likely to be associated with a 0.1 standard-deviation higher *MSRating*. In columns (5) and (6), we find qualitatively similar results when controlling *PastMSRating*, suggesting that not only are informed funds more likely to have a higher *MSRating*, but also have a greater likelihood to receive an upgrade in *MSRating*.

5. Can Investors recognize *Textual Fund Information*?

5.1. Fund flows

Since the predictions by the neural network model purely rely on textual information contained in the shareholder letters which target fund investors, a natural question is that whether letters written by informed funds attract more flows than those written by uninformed funds. A diagnostic test is thus to relate future fund flows to *Textual Fund Information* in the cross section. Table 4 reports the results from the following regression at the filing level, indexed by fund(i)-filing(j)-date(t), with year and style fixed effects:

$$Flow_{i,j,t} = \beta Textual\ Fund\ Information_{\ i,j,t} + \gamma FundChar_{i,j,t} + \alpha_{style} + \alpha_{year} + \epsilon_{i,j,t} \quad (3)$$

Although *Flow6m* is the primary variable that represents flows, we consider flows in the long term and short term by using *Flow12m*, *Flow24m*, *Flow3d*, and *Flow5d*, measured in various windows. For flows measured beyond the 6-month window, we also control contemporaneous *Alpha* to rule out the possibility that investors simply chase concurrent returns instead of making investment based on information from shareholder letters.

The first two columns of Table 4 Panel A show that investors recognize *Textual Fund Information* by putting more capital into informed funds than uninformed ones. The difference between flows into two groups of funds is 1.59% to 1.82% over a 6-month horizon post filing. It is plausible that an uninformed fund manager can exaggerate his ability in the shareholder letter to reduce the redemption risk. However, investors will observe his true ability to collect private information in the long run, and we should not observe inflows to his fund in the long term. The last four columns of Table 4 Panel A report that funds predicted to perform better receive greater flows than those predicted to underperform beyond a semi-annual period and in up to two years after investors receive shareholder letters. The economic magnitude of the difference in flows between two groups is 2.84% to 3.00%

in the next year and 4.49% to 4.63% in the next two years. The results echo an incentive for informed funds that they want to reduce capital constraint to achieve their long-term investment strategy, even if they may not outperform in the short term.

[Insert Table 4 Here]

A presumable assumption for informed funds to attract flows is that investors pay attention to shareholder letters. Recent studies show that investment companies use Form N-CSR and 13-F to make investment decisions (Chen, Cohen, Gurun, Lou, and Malloy, 2020; Crane, Crotty, and Umar, 2020; Cao, Du, Yang, and Zhang, 2021). We zoom into the short window around the dates on which funds send their letters to investors and analyze the timing of flows. We posit that informed funds identified by the neural network model expect higher flows right after shareholder letters are accessible to investors. To test this hypothesis, we use daily flow data from the Trimtabs database that has been used previously in Greene and Hodges (2002), Kaniel and Parham (2017), and Agarwal, Jiang, and Wen (2020).

Results in Table 4 Panel B confirm the conjecture. Informed funds attract 11.5 to 13.5 bps more capital flows during days [0, 3] and 23.4 to 24.5 bps more during days [0, 5] where day 0 indexes the date on which funds file shareholder letters. We control style fixed effects to minimize the possibility that investors may allocate their capital into funds differentially because of various investment objectives. The results hold with year and style fixed effects. The economic magnitude of the difference in flows within five days is equivalent to a dollar value of $4.48 million for an average fund.

5.2. When are informed managers rewarded with greater capital flows?

One of the essential incentives for managers arises from the compensation scheme for managing funds. Although managers can choose management fees at different levels, total net assets (or capital flows) play an equally important role in determining their compensation. If managers endeavor to write enlightening shareholder letters in order to signal their skill

and attract capital flows, we should observe their success in doing so if fund investors indeed read their letters and make investment decisions accordingly.

It is challenging to discover whether investors read shareholder letters since we cannot track the identities of investors. To overcome the obstacle, we use the number of visits to a shareholder letter on the SEC EDGAR website to measure the investor attention to the letter, which proxies for the readership on the shareholder letter. We expect to observe more responsive capital flows to shareholder letters by informed funds when there exists a greater readership on their letters. On the other fund, flows are indifferent between informed funds and uninformed funds if the investor attention to shareholder letters is lower. We reexamine equation (3) for two subsamples that are partitioned based on the investor attention.

[Insert Table 5 Here]

Results in Table 5 provide supportive evidence that informed managers are successful in attracting more capital flows than uninformed managers when investors take their letters into consideration in making investment decisions (i.e., investor attention is high). Such results further validate that investors who invest in informed funds make decisions based on the informational contents in shareholder letters rather than other concurrent sources.

6. Informational Contents of Shareholder Letters

6.1. Topics by informed funds

Section 4 confirms that the neural network model designed to select informed funds achieves the goal and shows that selected funds produce higher future performance and are more likely to have higher Morningstar ratings as well as receive an upgrade in the ratings. Section 5 further finds investors recognize the skill derived from qualitative disclosure by putting greater flows within 3-day window and in up to two years into informed funds. In this section, we zoom into the model to understand the source of the predictive power.

We first explore choices of topics discussed by informed funds. We exploit unsupervised LDA, developed by Blei, Ng, and Jordan (2003), to classify shareholder letters into four categories. For each category, we select the top five keywords to identify the potential topics and assign a label based on these keywords.

[Insert Table 6 Here]

Table 6 shows the list of topics discussed by informed funds. We observe that the first group of funds tend to discuss sectors and are likely to be specialized in certain sectors such as energy. The second category of funds focus on the risk taking of the portfolio and the third group of funds incline to offer insights on the big picture of the financial market and the economy. Although our sample consists of domestic equity funds, the conventional selection criteria in Kacperczyk, Sialm, and Zheng (2008) do not rule out the possibility that funds that invest primarily in the equity market can hold a small proportion of assets in bonds. Interestingly, informed funds in the last category frequently discuss the bond yield, and the interest rate, which depends on Fed policies. It is indicative that equity funds that hold mixed financial securities are likely to perform well since these managers have the ability to gather private information outside the equity market.

6.2. Who includes quality information in shareholder letters?

The underlying assumption for our neural network model to work well is that managers provide informative discussions in shareholder reports. The results in Section 4 validate such an assumption. However, given that disclosure is costly because competitors can copycat the investment strategy of disclosing companies (Frank, Poterba, Shackelford, and Shoven, 2004; Phillips, Pukthuanthong, and Rau, 2014; Cao, Du, Yang, and Zhang, 2021), it begs another question that to what extent disclosing quality information is optimal.

We explore what type of funds are capable of writing more informative textual contents. We thus relate the accuracy of prediction to fund characteristics and consider the following

Logit regressions:

$$AccuratePrediction_{i,j,t} = \gamma FundChar_{i,j,t} + \alpha_{year} + \epsilon_{i,j,t} \tag{4}$$

Table 7 shows that our neural network model generates greater accuracy in funds engaging in higher risk taking, measured as volatility of past fund returns. For funds with higher risk taking, it is especially hard to distinguish their skill from luck because of volatile historical returns. Therefore, our model adds unique value on evaluating these opaque funds and can potentially provide insights to fund investors.

[Insert Table 7 Here]

The greater predictive power in high risk-taking funds can arise from a more detailed and informative shareholder letter. To examine such a possibility, we conduct cross-sectional analysis on high risk-taking funds and low risk-taking funds by replicating equations (1), (2) and (3).

Table 8 Panel A shows that among "High Risk Funds," informed funds outperform uninformed funds by 1.32% to 1.40% annually on a risk-adjusted basis. The outperformance is much weaker or insignificant among "Low Risk Funds." Table 8 Panel B reports that among funds with high risk-taking, informed ones are more likely to receive an upgrade in Morningstar ratings compared with their uninformed counterparties. Such a result is nonexistent among funds with low risk taking.

[Insert Table 8 Here]

In Table 8 Panel C, High Risk Funds are more likely to attract capital flows if they are evaluated to possess the skill. The incremental inflows are significant regardless of the horizons over which flows are measured (from 3-day to 24-month). On the other hand, Low Risk Funds on average experience the same amount of flows, and there is no difference between informed funds and uninformed funds.

In sum, because our neural network model relies entirely on the textual contents in shareholder letters, the results in Table 8 supports the view that informed managers that take greater risk are more likely to write an informative letter to investors in order to distinguish themselves from uninformed managers.

6.3. Voluntary disclosure and fund characteristics

To further understand the incentives for managers to disclose their privileged information, we compare funds in our sample with funds that do not include a shareholder letter in Form N-CSR[12] and examine whether the decision to include a letter is related to any fund characteristics. Because there are only two outcomes (i.e., to include or not to include a letter), we consider a determinant model with Logit regressions:

$$ShareholderLetter_{i,j,t} = \gamma FundChar_{i,j,t} + \alpha_{year} + \epsilon_{i,j,t} \tag{5}$$

Table 9 represents the relations between fund characteristics and the likelihood of including a shareholder letter in Form N-CSR. Funds with higher risk taking are more likely to not only write informative shareholder letters (Section 6.2 and Tables 7 and 8) but also include qualitative disclosure in their filings in the first place. There is additional evidence that funds with higher turnover rates tend to write shareholder letters since it is difficult for investors to disentangle their true skill from frequent portfolio changes.

[Insert Table 9 Here]

The likelihood to include a shareholder letter is positively associated with 12b-1 fees, indicating funds that spend remarkable expenses on marketing and distributions are likely to utilize shareholder letters as an efficient tool to communicate with investors. Older funds

[12]Some funds choose not to provide any subjective discussion. Instead, they report quantitative information readily available from other data sources such as performance in the last period or historical performance since a fund's inception. We consider these funds as funds that do not include a shareholder letter because of the lack of any qualitative information.

and funds with longer manager tenure are more likely to write shareholder letters because managers may have less career concern and, therefore, greater likelihood to offer voluntary discussions spanning various topics that intrigue potential investors, consistent with the notion that disclosing valuable information that bears out in the future signals managerial skill and can help the manager in the future labor market (Stern and James, 2016). Because future labor market opportunities include both the external jobs and internal promotions, the results also highlight the incentive to disclose true information since any falsified information can be easily verified by the current employers and thus detrimental to plausible internal promotions.

7. Concluding Remarks

This paper creates an innovative mutual fund private information measure, *Textual Fund Information*, by implementing NLP and neural network models on textual information contained in mutual fund shareholder letters. The neural network model that we design to predict future fund performance is more competent than traditional bag-of-words approaches in textual analysis. It is successful in identifying informed funds that deliver better abnormal returns and receive higher Morningstar ratings. We open the black box machine learning by compiling a list of topics discussed by informed funds, including sector specialization, portfolio risk taking, big picture of the financial market, and mixed strategies across assets. The informational contents on these topics contribute to the predictive power of *Textual Fund Information*.

Furthermore, funds with *Textual Fund Information* attract greater capital flows in both short-term and long-term than those without, suggesting investors are capable of recognizing such skill; the results are more pronounced when more investors take letters into account in making investment decisions.

Our model produces greater accuracy for funds with a riskier portfolio, which are generally viewed as opaque funds because their true skill is hard to distinguish from luck. More-

over, these funds are more likely to write informative letters than funds that take a lower risk. Funds with higher risk taking, higher portfolio turnover rates, greater marketing expenses, longer manager tenure, and older funds are more likely to include qualitative disclosure in Form N-CSR because they view shareholder letters as necessary tools to communicate with investors.

Our applications of NLP models and neural networks capture higher-order syntactic interactions among words and sentences and convey unstructured textual data into numerical information that becomes accessible to researchers. Our study contributes to the growing literature on applying machine learning in exploring finance-related questions that are under debate or still lacking an answer. While we rely entirely on textual data to build our model, it is worthwhile to expect that future research can complement textual data with traditional numerical data to develop promising applications.

Appendix A: Definitions of Variables

Variable	Definition
Textual Fund Information	We use BERT to convert a fund's textual shareholder letter to numerical parameters and then build a neural network model to predict a label on the future performance of a fund. Each year, funds are predicted with labels of 2, 1, and 0. *Textual Fund Information* is an indicator variable equal to one if a fund has a label of 2 and zero if it has a label of 0.
Alpha	The Fama-French-Carhart four-factor alpha using daily returns during days [0, 180] for a fund's filing on day 0, expressed in percentage points.
Flow6m (*Flow12m*, *Flow 24m*)	The future 6-month flow for the filing of fund i in month t, expressed in percentage points. $Flow6m = (TNA_{i,t+5} + TNA_{i,t-1} \times R_{i,[t-1,t+5]})/TNA_{i,t-1}$ *Flow12m* and *Flow24m* are calculated analogously.
Flow3d (*Flow5d*)	The flow during days [0, 3] ([0, 5]) for a fund's filing on day 0, expressed in percentage points.
MSRating	Fund rating calculated by Morningstar in month $t + 5$where t is the filing month.
AccuratePrediction	An indicator variable equal to one if a fund with *Textual Fund Information* equal to one (zero) is in the ex-post top (bottom) tercile of *Alpha*, and zero otherwise.
ShareholderLetter	An indicator variable equal to one if a fund includes a qualitative shareholder letter in N-CSR reports, and zero otherwise.
PastRisk	The return volatility during days [-180, -1] for a fund's filing on day 0.
PastAlpha	The Fama-French-Carhart four-factor alpha using daily returns during days [-180, -1] for a fund's filing on day 0, expressed in percentage points.
PastFlow	The past 6-month flow during months [t – 6, t – 1] for a fund's filing in month t.
Log(TNA)	The natural logarithm of a fund's TNA ($mm) in month t – 1.
Expense	The most recent expense ratio prior to filing month t.
Turnover	The most recent turnover ratio prior to filing month t.
Log(Age)	The natural logarithm of a fund's age.
LM_Negative	The number of Loughran-McDonald (LM) finance-related negative words in a shareholder letter divided by the length (i.e., total number of words) of the letter, expressed in percentage points.
LM_Positive	The number of LM finance-related positive words in a shareholder letter divided by the length of the letter, expressed in percentage points.
DocumentSize	The natural logarithm of the number of words in a shareholder letter.
TextDiversity	Carroll (1964) corrected type-token ratio, defined as the ratio of the number of distinct words in a shareholder letter to the square root of the total number of words in the letter.
HoldingReturn	The raw return on a fund's portfolio holdings during months [t, t + 5] for a fund's filing in month t. .
DGTWReturn	The DGTW benchmark-adjusted return on a fund's portfolio holdings during months [t, t + 5] for a fund's filing in month t. .

(continued)

Variable	**Definition**
PastMSRating	Fund rating calculated by Morningstar in month $t - 1$ where t is the filing month.
PastReturnGap	The average return gap during months $[t - 6, t - 1]$ for a fund's filing in month t. The calculation of the return gap follows Kacperczyk, Sialm, and Zheng (2008).
Tenure	The number of years since a portfolio manager is hired. If there are multiple managers for a fund, the longest tenure is used.
12b1Fee	The most recent 12b-1 fees prior to filing month t.
MgmtFee	The most recent management fees prior to filing month t.

Appendix B: Details of Bidirectional Encoder Representations from Transformers (BERT)

Transformers are designed to handle tasks such as text summarization, which finds the most informative sentences in a document, using sequential data. BERT build on Transformers and learn contextual relations among words in a text. Unlike traditional NLP models that input the text sequentially (from left to right or from right to left), BERT read entire sequence of words at the same time and thus are able to learn the context of a word using all its surroundings.

The embedding representation of a word (i.e., token), denoted by E, is input from the embedding layer. Take the first token in Figure A1 as an example, E_1 is the embedding representation and T_1 is the final output of the first token. T_{rm} are the representations of the same token in each intermediate layer. The base BERT model that we use in this paper have 12 intermediate layers.

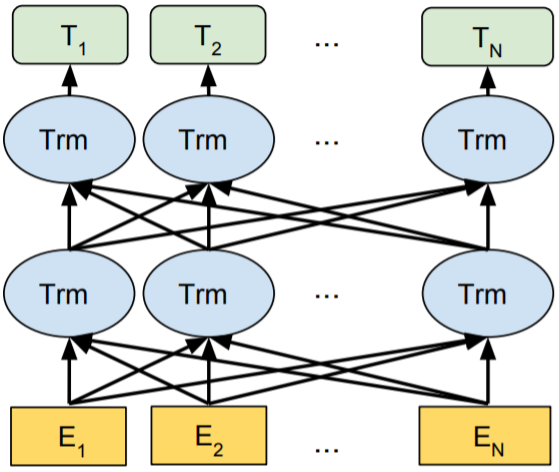

Figure A1: BERT architectures (source: reprinted from Devlin et al., 2019)

For each token, its input representation is the sum of the token embedding, the segment embedding, and the position embedding. As Figure A2 shows, the embedding representation of a word not only captures the meaning of the word, but also reads its sequence and relations

with the surrounding context. Special tokens [CLS] and [SEP] are used to separate sentences.

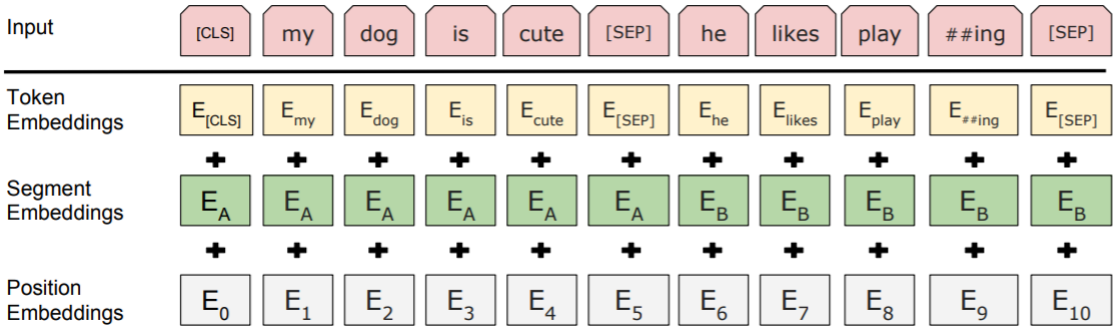

Figure A2: BERT input representation (source: reprinted from Devlin et al., 2019)

When using BERT to read sentences, a user is required to add a special token [CLS] as the first token of every sentence. This special token in the final hidden state is the aggregate sequence representation, denoted as a vector $C \in \mathbb{R}^{\mathrm{H}}$, where H is the hidden size and equals to 768 in the base BERT model. The final hidden state therefore corresponds to C, T_1, T_2, $\ldots T_N$. We use the vector C as the features (i.e., aggregate representation) extracted from a sentence via BERT.

References

Admati, A., and P. Pfleiderer. 2000. Forcing firms to talk: Financial disclosure regulation and externalities. *Review of Financial Studies* 13: 479–519.

Agarwal, Vikas, Lei Jiang, and Quan Wen, 2020, Why Do Mutual Funds Hold Lottery Stocks? *Journal of Financial and Quantitative Analysis* forthcoming.

Ali, Ashiq, Sandy Klasa, and Eric Yeung, 2014, Industry concentration and corporate disclosure policy, *Journal of Accounting and Economics* 58: 240–264.

Amihud, Yakov, and Ruslan Goyenko, 2013, Mutual fund's R2 as predictor of performance, *Review of Financial Studies* 26, 667–694.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2019, What do mutual fund investors really care about? Working paper, Ohio State University, University of Utah, University of Arizona, and University of Washington.

Bernard, Darren, 2016, Is the risk of product market predation a cost of disclosure? *Journal of Accounting and Economics* 62: 305-â325.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan, 2003, Latent dirichlet allocation, *Journal of Machine Learning Research* 3, 993–1022.

Brands, Simone, Stephen J. Brown, and David R. Gallagher, 2005, Portfolio concentration and investment manager performance, *International Review of Finance* 5,149–174.

Cao, Sean, Kai Du, Baozhong Yang, and Alan L. Zhang, 2021, Copycat skills and disclosure costs: Evidence from peer companies' digital footprints, *Journal of Accounting Research*, forthcoming.

Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L. Zhang, 2021, How to talk when a machine is listening: Corporate disclosure in the age of AI, Working paper, Columbia University and Georgia State University.

Cao, Sean, Guang Ma, Jennifer Wu Tucker, and Chi Wan 2018, Technological peer pressure and product disclosure, *The Accounting Review* 93: 95–126.

Carhart, Mark M, 1997, On persistence in mutual fund performance, *Journal of Finance* 52, 57–82.

Carroll, John B., *Language and Thought*, Prentice Hall, Hoboken, NJ.

Chen, Danqi, and Christopher Manning, 2014, A fast and accurate Dependency Parser using Neural Networks. *Proceedings of EMNLP*.

Chen, Huaizhi, Lauren Cohen, Umit Gurun, Dong Lou, and Christopher Malloy, 2020, IQ from IP: Simplifying search in portfolio choice, *Journal of Financial Economics* forthcoming.

Cheng, Si, Ruichang Lu, and Xiaojun Zhang, 2021, What Should Investors Care About? Mutual fund ratings by analysts vs. machine learning technique, Working paper, Chinese University of Hong Kong and Peking University.

Crane, Alan D. and Kevin Crotty and Tarik Umar, 2020, Public and private information: complements or substitutes? Working paper, Rice University.

Cremers, K. J. Martijn, and Antti Petajisto, 2009, How active is your fund manager? A new measure that predicts performance, *Review of Financial Studies* 22, 3329–3365.

Cremers, K. J. Martijn, Miguel A. Ferreira, Pedro Matos, and Laura Starks, 2016, Indexing and active fund management: International evidence, *Journal of Financial Economics* 120, 539–560.

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristicâbased benchmarks, *Journal of Finance* 52, 1035–1058.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2019, Bert: Pre-training of deep bidirectional transformers for language understanding, Working paper, Google AI Language.

Diamond, D., and R. Verrecchia. 1991. Disclosure, liquidity, and the cost of capital, *Journal of Finance* 46, 1325–1359.

Evans, Richard B., and Yang Sun, 2021, Models or stars: The role of asset pricing models and heuristics in investor risk adjustment, *Review of Financial Studies* 34, 67–107.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Finance* 33, 3–56.

Ferson, Wayne, and Haitao Mo, 2016, Performance measurement with selectivity, market and volatility timing, *Journal of Financial Economics* 121, 93–110.

Frank, Mary Margaret, James M. Poterba, Douglas A. Shackelford, and John B. Shoven, 2004, Copycat funds: Information disclosure regulation and the returns to active management in the mutual fund industry, *Journal of Law and Economics* 47, 515–541.

Garcia, Diego, Xiaowen Hu and Maximilian Rohrer, 2020, The colour of finance words, Working paper, University of Colorado at Boulder and Norwegian School of Economics.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–74.

Greene, Jason T., and Charles W. Hodges, 2002, The dilution impact of daily fund flows on open-end mutual funds, *Journal of Financial Economics* 65, 131–158.

Hillert, Alexander, Alexandra Niessen-Ruenzi, and Stefan Ruenzi, 2016, Mutual fund shareholder letters: Flows, performance, and managerial behavior, Working paper, Goethe University Frankfurt and University of Mannheim.

Kacperczyk, Marcin, and Amit Seru, 2007, Fund manager use of public information: New evidence on managerial skills, *Journal of Finance* 62, 485–528.

Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng, 2005, On the industry concentration of actively managed equity mutual funds, *Journal of Finance* 60, 1983–2011.

Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng, 2008, Unobserved actions of mutual funds, *Review of Financial Studies* 21, 2379–2416.

Kaniel, Ron, and Robert Parham, 2017, WSJ Category Kings–The impact of media attention on consumer and mutual fund investment decisions, *Journal of Financial Economics* 123, 337–356.

Ke, Zheng, Bryan T. Kelly, and Dacheng Xiu, 2020, Predicting returns with text data, Working paper, Harvard University, Yale University and University of Chicago.

Kelly, Bryan T., Asaf Manela, and Alan Moreira, 2019, Text selection, Working paper, Yale University, Washington University in St. Louis and University of Rochester.

Li, Yinghua, Yupeng Lin, and Liandong Zhang, 2018, Trade secrets law and corporate disclosure: Causal evidence on the proprietary cost hypothesis, *Journal of Accounting Research* 56: 265–308.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *Journal of Finance* 66, 35–65.

Loughran, Tim, and Bill McDonald, 2014, Measuring readability in financial disclosures, *Journal of Finance* 69, 1643–1671.

Loughran, Tim, and Bill McDonald, 2016, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* 54, 1187–1230.

Pappagari, Raghavendra, Piotr Zelasko, Jesus Villalba, Yishay Carmiel, and Najim Dehak, 2019, Hierarchical transformers for long document classification, *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 838–844.

Phillips, Blake, Kuntara Pukthuanthong, and P. Raghavendra Rau, 2014, Detecting superior mutual fund managers: Evidence from copycats, *Review of Asset Pricing Studies* 4, 286–321.

Sirri, Erik R., and Peter Tufano, 1998, Costly search and mutual fund flows, *Journal of Finance* 53, 1589–1622.

Stern, Ithai, and Sharon D. James, 2016, Whom are you promoting? Positive voluntary public disclosures and executive turnover, *Strategic Management Journal* 37, 1413–1430.

Table 1: Summary Statistics

This table provides summary statistics. Fund characteristics and filing level variables are based on the sample of funds and their shareholder letters from 2015 to 2018. Variables are defined in Appendix A.

Variables	(1) Mean	(2) Median	(3) Std	(4) P25	(5) P75	(6) N
Alpha	-1.45	-1.14	8.38	-4.96	2.76	6,002
Flow6m	-1.25	-4.74	31.96	-9.59	0.43	5,896
Flow12m	-2.51	-9.74	46.09	-18.44	0.36	5,806
Flow24m	-3.65	-20.36	70.77	-35.35	-0.75	4,144
Flow3d	-0.10	-0.05	2.16	-0.18	0.04	2,899
Flow5d	-0.15	-0.09	2.64	-0.28	0.06	2,899
MSRating	3.05	3.00	0.94	2.33	3.75	5,267
PastRisk	11.29	11.86	6.08	6.77	15.64	6,000
PastAlpha	-1.49	-1.10	7.87	-4.96	2.72	6,000
PastFlow	-0.03	-4.44	36.29	-9.21	0.91	5,952
Log(TNA)	5.86	5.89	1.98	4.54	7.22	6,000
Expense	1.02	1.00	0.40	0.78	1.23	5,839
Turnover	73.70	44.00	103.30	23.00	80.00	5,839
Log(Age)	2.94	3.07	0.74	2.80	3.32	6,000
PastMSRating	3.05	3.00	0.95	2.33	3.80	5,232
LM_Negative	2.06	1.92	0.91	1.48	2.68	6,002
LM_Positive	1.84	1.78	1.23	0.94	2.65	6,002
DocumentSize	5.60	5.65	0.79	5.11	6.19	6,002
TextDiversity	8.49	8.63	2.22	7.23	9.92	6,002
HoldingReturn	0.00	0.00	0.66	-0.27	0.29	3,312
DGTWReturn	0.82	1.06	1.59	-0.03	1.88	3,514

Table 2: Future Performance and *Textual Fund Information*

This table reports the regressions of future abnormal returns on *Textual Fund Information* and fund characteristics. Variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by fund. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable			*Alpha*			
Textual Fund Information	0.848***	0.851***	0.864***	0.785***	0.881***	1.008**
	(3.91)	(3.94)	(3.66)	(2.92)	(2.79)	(2.29)
PastRisk	-0.275***	-0.274***	-0.256***	-0.252***	-0.278***	-0.226***
	(-11.00)	(-10.51)	(-8.03)	(-4.87)	(-4.54)	(-2.91)
PastAlpha	-0.050**	-0.055**	-0.089***	-0.179***	-0.205***	-0.158***
	(-2.31)	(-2.52)	(-4.01)	(-7.57)	(-7.43)	(-4.19)
PastFlow	-0.000	-0.001	-0.003	-0.002	-0.005	-0.003
	(-0.17)	(-0.26)	(-1.07)	(-0.38)	(-0.80)	(-0.61)
Log(TNA)	-0.109*	-0.109*	-0.140	-4.749***	-5.080***	-3.327***
	(-1.69)	(-1.67)	(-1.51)	(-8.90)	(-7.37)	(-3.88)
Expense	-1.710***	-1.721***	-1.598***	-6.483***	-5.173**	-2.492
	(-4.19)	(-4.23)	(-2.70)	(-3.00)	(-2.07)	(-0.76)
Turnover	-0.002	-0.002	-0.002	0.002	0.003	-0.006
	(-1.15)	(-1.29)	(-0.74)	(0.63)	(0.67)	(-0.77)
Log(Age)	0.077	0.089	-0.171	6.227***	5.947***	3.522
	(0.46)	(0.53)	(-0.73)	(4.24)	(3.36)	(1.59)
LM_Negative					0.021	0.233
					(0.11)	(0.90)
LM_Positive					0.634***	0.553**
					(3.33)	(2.28)
DocumentSize					1.882**	3.274***
					(2.27)	(2.65)
TextDiversity					-0.671**	-0.984**
					(-2.24)	(-2.32)
HoldingReturn						1.218***
						(3.97)
DGTWReturn						2.968***
						(5.37)
Observations	5,803	5,803	5,766	5,505	5,156	2,703
R-squared	0.057	0.064	0.164	0.326	0.361	0.461
Manager FE	No	No	No	No	Yes	Yes
Fund FE	No	No	No	Yes	Yes	Yes
Family FE	No	No	Yes	No	No	No
Year FE	No	Yes	Yes	Yes	Yes	Yes

Table 3: Future Morningstar Rating and *Textual Fund Information*

This table reports the regressions of future Morningstar rating on *Textual Fund Information* and fund characteristics. Variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by fund. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable			*MorningstarRating*			
Textual Fund Information	0.096***	0.099***	0.104***	0.038**	0.053***	0.048***
	(3.42)	(3.52)	(4.25)	(2.25)	(3.88)	(3.26)
PastRisk	0.001	0.002	-0.005	0.004	-0.004**	-0.003
	(0.35)	(0.41)	(-1.02)	(1.11)	(-2.11)	(-1.13)
PastAlpha	0.020***	0.021***	0.015***	0.007***	0.003***	0.003***
	(11.58)	(11.55)	(9.19)	(5.93)	(3.22)	(3.27)
PastFlow	0.003***	0.003***	0.002**	0.001*	0.000*	0.001**
	(2.99)	(2.99)	(2.34)	(1.94)	(1.86)	(2.09)
Log(TNA)	0.156***	0.156***	0.186***	0.021	0.019***	-0.066**
	(11.17)	(11.13)	(10.32)	(0.53)	(2.95)	(-2.06)
Expense	-0.301***	-0.298***	0.055	0.134	0.056	0.038
	(-4.39)	(-4.28)	(0.50)	(0.62)	(1.55)	(0.24)
Turnover	-0.000	-0.000	-0.000	0.000	0.000	0.000
	(-0.67)	(-0.66)	(-0.25)	(0.99)	(0.41)	(1.53)
Log(Age)	-0.063	-0.066	-0.152***	-0.166	-0.023	-0.016
	(-1.50)	(-1.55)	(-2.84)	(-0.55)	(-1.00)	(-0.06)
PastMSRating					0.795***	0.449***
					(71.38)	(20.32)
Observations	5,112	5,112	5,077	4,876	5,032	4,832
R-squared	0.193	0.194	0.461	0.822	0.795	0.857
Fund FE	No	No	No	Yes	No	Yes
Family FE	No	No	Yes	No	Yes	No
Year FE	No	Yes	Yes	Yes	Yes	Yes

Table 4: Future Flows and *Textual Fund Information*

This table reports the regressions of future monthly flows and daily flows on *Textual Fund Information* and fund characteristics. Variables are defined in Appendix A.The t-statistics, in parentheses, are based on standard errors clustered by fund. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: Monthly Flows

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	*Flow6m*		*Flow12m*		*Flow24m*	
Textual Fund Information	1.587*	1.819**	2.844**	2.992**	4.489*	4.634**
	(1.93)	(2.23)	(2.30)	(2.41)	(1.94)	(2.01)
PastRisk	0.283***	0.358***	0.646***	0.737***	1.165***	1.480***
	(3.67)	(2.95)	(4.31)	(3.46)	(3.31)	(3.44)
PastAlpha	0.285***	0.330***	0.489***	0.562***	0.919***	1.024***
	(4.81)	(5.72)	(4.98)	(6.07)	(4.96)	(5.89)
PastFlow	0.051*	0.040	0.160***	0.148***	0.287***	0.280***
	(1.71)	(1.18)	(3.46)	(3.07)	(3.56)	(3.48)
Log(TNA)	-1.805***	-1.865***	-3.213***	-3.305***	-5.099***	-5.340***
	(-5.07)	(-5.30)	(-5.72)	(-5.81)	(-4.55)	(-4.69)
Expense	-7.189***	-7.430***	-11.042***	-10.903***	-15.748***	-14.278***
	(-4.92)	(-4.99)	(-4.70)	(-4.59)	(-3.48)	(-3.16)
Turnover	0.016***	0.014**	0.023**	0.021**	0.004	0.007
	(2.74)	(2.54)	(2.55)	(2.12)	(0.32)	(0.49)
Log(Age)	-3.459***	-3.665***	-6.684***	-7.129***	-17.570***	-17.644***
	(-3.40)	(-3.61)	(-3.90)	(-4.02)	(-5.66)	(-5.65)
Alpha			0.756***	0.751***	1.167***	1.165***
			(9.31)	(9.26)	(6.24)	(6.08)
Observations	5,687	5,664	5,624	5,601	4,063	4,047
R-squared	0.041	0.055	0.093	0.110	0.135	0.161
Style FE	No	Yes	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Daily Flows

	(1)	(2)	(3)	(4)
Dependent Variable	*Flow3d*		*Flow5d*	
Textual Fund Information	0.115*	0.135**	0.245**	0.234***
	(1.73)	(2.12)	(2.36)	(2.64)
PastRisk	-0.001	0.017	-0.000	0.016
	(-0.11)	(1.54)	(-0.03)	(1.37)
PastAlpha	0.019**	0.020**	0.021**	0.019**
	(2.00)	(2.30)	(2.45)	(2.41)
PastFlow	0.003*	0.002	0.001	0.001
	(1.73)	(1.53)	(0.44)	(0.26)
Log(TNA)	-0.006	-0.004	-0.022	-0.016
	(-0.31)	(-0.18)	(-0.93)	(-0.62)
Expense	0.043	0.025	-0.054	-0.143
	(0.38)	(0.24)	(-0.37)	(-0.89)
Turnover	-0.003	-0.002	-0.000	-0.000
	(-1.22)	(-1.49)	(-0.14)	(-0.00)
Log(Age)	-0.004***	-0.003**	-0.006***	-0.005**
	(-2.84)	(-2.08)	(-3.31)	(-2.33)
Observations	2,789	2,784	2,789	2,784
R-squared	0.026	0.056	0.011	0.044
Style FE	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes

Table 5: Future Flows and Investor Attention to Shareholder Letter

This table reports the regressions of future flows on *Textual Fund Information* and fund characteristics for subsamples partitioned based on the investor attention, measured by the number of downloads of the fund's shareholder letter on the SEC EDGAR website. Funds are sorted into two groups by median. The group above (below) median is labeled as "High (Low) Investor Attention Funds." Variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by fund. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Subgroups	High Investor Attention Funds				Low Investor Attention Funds			
Dependent Variable	*Flow6m*	*Flow12m*	*Flow3d*	*Flow5d*	*Flow6m*	*Flow12m*	*Flow3d*	*Flow5d*
Textual Fund Information	3.914**	5.380**	0.239*	0.376**	0.702	2.139	0.018	0.079
	(2.04)	(2.32)	(1.73)	(2.09)	(0.59)	(1.06)	(0.22)	(0.79)
PastRisk	0.447	0.886**	0.024	0.034	0.216	0.967**	0.014	0.023
	(1.61)	(2.46)	(1.24)	(1.35)	(1.23)	(2.27)	(0.54)	(0.92)
PastAlpha	0.427***	0.816***	0.012	0.003	0.239***	0.612***	0.031*	0.034**
	(3.15)	(4.83)	(1.23)	(0.19)	(3.15)	(3.34)	(1.85)	(2.21)
PastFlow	-0.018	0.070	0.001	-0.003	0.093**	0.220***	0.005*	0.006**
	(-0.32)	(1.00)	(0.60)	(-0.92)	(2.44)	(2.86)	(1.72)	(2.08)
Log(TNA)	-3.630***	-5.423***	0.016	-0.017	-0.481	-1.369	0.063	0.088
	(-3.88)	(-4.12)	(0.27)	(-0.25)	(-1.12)	(-1.59)	(1.02)	(1.57)
Expense	-12.595***	-18.249***	0.166	-0.252	-3.898***	-6.007*	0.346	0.238
	(-3.35)	(-3.64)	(0.87)	(-0.70)	(-2.59)	(-1.92)	(0.98)	(0.71)
Turnover	0.019	0.024	-0.002**	0.002	0.011	0.026	-0.006	-0.003
	(1.56)	(1.14)	(-1.99)	(0.91)	(1.15)	(1.40)	(-1.07)	(-0.51)
Log(Age)	-5.066**	-9.482***	-0.291***	-0.316**	-3.695***	-7.706***	-0.121	-0.157*
	(-2.49)	(-3.33)	(-2.64)	(-2.21)	(-2.88)	(-3.23)	(-1.31)	(-1.84)
Alpha		0.861***				0.773***		
		(5.96)				(5.62)		
Observations	1,704	1,689	873	873	1,769	1,756	879	879
R-squared	0.086	0.153	0.076	0.055	0.070	0.122	0.174	0.164
Style FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 6: Topics Discussed by Funds with *Textual Fund Information*

This table reports the top five keywords for each topic classified by LDA.

Category label	Top keywords				
Sector specialization	sector	information	relative	benchmark	energy
Portfolio risk taking	portfolio	value	believe	asset	risk
Bid picture of financial market	financial	global	economy	volatility	policy
Strategies in mixed assets	rate	bond	equity	yield	fed

Table 7: The Accuracy of Predictions and Fund Characteristics

This table reports relations between the accuracy of predictions and fund characteristics using Logit regressions. Variables are defined in Appendix A. The z-statistics, in parentheses, are based on standard errors clustered by fund. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)
Dependent Variable		*AccuratePrediction*		
PastRisk	0.019***	0.020***	0.030***	0.034***
	(3.63)	(3.52)	(3.38)	(3.38)
PastAlpha	-0.001	-0.001	-0.003	-0.003
	(-0.16)	(-0.18)	(-0.55)	(-0.55)
PastFlow	0.000	0.000	0.001	0.001
	(0.22)	(0.22)	(1.15)	(1.18)
Log(TNA)	-0.053***	-0.053***	-0.046	-0.045
	(-2.83)	(-2.84)	(-1.46)	(-1.43)
Expense	0.187**	0.184**	0.172	0.177
	(2.08)	(2.03)	(1.24)	(1.26)
Turnover	0.000	0.000	0.000	0.000
	(0.32)	(0.31)	(0.40)	(0.29)
Log(Age)	0.047	0.047	0.004	0.004
	(1.09)	(1.09)	(0.05)	(0.05)
PastMSRating			0.027	0.027
			(0.59)	(0.59)
LM_Negative			0.067	0.067
			(1.46)	(1.47)
LM_Positive			-0.079**	-0.072*
			(-2.12)	(-1.91)
DocumentSize			0.155	0.163
			(1.12)	(1.17)
TextDiversity			-0.011	-0.015
			(-0.23)	(-0.31)
HoldingReturn			-0.035	-0.044
			(-1.21)	(-1.25)
DGTWReturn			-0.008	-0.002
			(-0.12)	(-0.03)
Observations	5,803	5,803	2,880	2,880
Year FE	No	Yes	No	Yes
Pseudo R-squared	0.007	0.007	0.011	0.012

Table 8: Future Performance, Flows, Morningstar Ratings and *Textual Fund Information*: Funds Risk Taking

This table reports the regressions of future abnormal returns, flows, and Morningstar ratings on *Textual Fund Information* and fund characteristics for subsamples partitioned by the past risk taking of funds. Funds are sorted into two groups by median based on *PastRisk*. The group above the median is labeled as "High Risk Funds" and the group below the median is labeled as "Low Risk Funds." Variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by fund. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

Panel A: Future performance

	(1)	(2)	(3)	(4)
Subgroups	High Risk Funds		Low Risk Funds	
Dependent Variable	*Alpha*			
Textual Fund Information	1.400***	1.324***	0.397*	0.039
	(4.09)	(2.89)	(1.75)	(0.14)
PastRisk	-0.297***	-0.177*	-0.225***	0.148
	(-3.87)	(-1.89)	(-5.78)	(1.28)
PastAlpha	-0.038	-0.241***	-0.128***	-0.305***
	(-1.47)	(-7.96)	(-4.32)	(-9.23)
PastFlow	-0.002	-0.000	0.000	-0.007
	(-0.62)	(-0.01)	(0.17)	(-1.34)
Log(TNA)	-0.267**	-5.462***	0.019	-3.171***
	(-2.27)	(-7.41)	(0.26)	(-4.90)
Expense	-2.630***	-7.489**	-1.048**	-2.418
	(-4.17)	(-2.32)	(-2.52)	(-1.09)
Turnover	-0.001	0.004	-0.002**	0.002
	(-0.53)	(0.52)	(-2.43)	(0.84)
Log(Age)	0.187	5.750***	0.214	4.013
	(0.73)	(2.69)	(1.08)	(1.54)
Observations	2,831	2,457	2,905	2,428
R-squared	0.036	0.389	0.076	0.422
Fund FE	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes

Panel B: Future Morningstar ratings

	(1)	(2)	(3)	(4)
Subgroups	High Risk Funds		Low Risk Funds	
Dependent Variable	*MorningstarRating*			
Textual Fund Information	0.074***	0.075***	0.021	-0.001
	(3.76)	(3.13)	(1.30)	(-0.03)
PastRisk	-0.006*	-0.003	0.003	-0.010
	(-1.77)	(-0.63)	(0.94)	(-1.12)
PastAlpha	0.004***	0.003**	0.005***	0.001
	(3.68)	(2.46)	(2.62)	(0.50)
PastFlow	0.000	0.000	0.001**	0.001
	(1.10)	(1.36)	(2.33)	(1.41)
Log(TNA)	0.010	-0.058	0.006	-0.026
	(1.33)	(-1.25)	(1.04)	(-0.54)
Expense	-0.047	-0.014	-0.002	-0.002
	(-1.42)	(-0.06)	(-0.07)	(-0.01)
Turnover	0.000	0.000	0.000	0.000
	(0.13)	(0.17)	(0.25)	(1.41)
Log(Age)	-0.010	-0.054	0.041**	0.407
	(-0.40)	(-0.13)	(2.04)	(0.78)
PastMSRating	0.832***	0.420***	0.868***	0.396***
	(67.66)	(13.01)	(85.29)	(10.45)
Observations	2,443	2,144	2,624	2,205
R-squared	0.749	0.860	0.792	0.884
Fund FE	No	Yes	No	Yes
Year FE	Yes	Yes	Yes	Yes

Panel C: Future flows

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Subgroups	High Risk Funds				Low Risk Funds			
Dependent Variable	*Flow6m*	*Flow12m*	*Flow3d*	*Flow5d*	*Flow6m*	*Flow12m*	*Flow3d*	*Flow5d*
Textual Fund Information	2.306*	4.291**	0.188**	0.315**	1.144	1.674	0.013	0.038
	(1.75)	(2.30)	(2.07)	(2.56)	(1.29)	(1.09)	(0.21)	(0.52)
PastRisk	0.174	0.801*	0.055	0.047	0.946***	1.611***	-0.021	-0.001
	(0.70)	(1.93)	(1.57)	(1.52)	(3.11)	(2.94)	(-1.15)	(-0.06)
PastAlpha	0.325***	0.585***	0.023*	0.023**	0.361***	0.710***	0.013***	0.011*
	(4.44)	(4.92)	(1.94)	(2.23)	(4.31)	(4.34)	(3.17)	(1.68)
PastFlow	0.031	0.159**	0.003	-0.000	0.045*	0.131***	0.001	0.002
	(0.58)	(2.26)	(1.20)	(-0.08)	(1.78)	(3.10)	(1.63)	(1.56)
Log(TNA)	-2.631***	-4.307***	0.004	0.004	-1.114***	-2.500***	0.008	0.011
	(-4.27)	(-4.49)	(0.13)	(0.12)	(-2.79)	(-3.71)	(0.46)	(0.45)
Expense	-10.265***	-16.332***	0.066	-0.001	-5.311***	-6.535**	0.025	-0.125
	(-4.49)	(-4.71)	(0.48)	(-0.00)	(-2.72)	(-2.00)	(0.21)	(-0.90)
Turnover	0.030***	0.036**	-0.004*	-0.001	0.000	0.009	0.000	0.001
	(2.96)	(2.11)	(-1.67)	(-0.22)	(0.04)	(0.83)	(0.05)	(1.42)
Log(Age)	-3.337**	-6.866***	-0.157**	-0.255***	-3.522***	-6.231**	-0.063	-0.088
	(-2.44)	(-3.11)	(-2.56)	(-2.73)	(-2.69)	(-2.41)	(-1.11)	(-1.48)
Alpha		0.751***				0.783***		
		(7.88)				(5.27)		
Observations	2,796	2,766	1,850	1,850	2,865	2,832	930	930
R-squared	0.066	0.124	0.066	0.051	0.066	0.113	0.082	0.071
Style FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 9: Determinants of Writing Shareholder Letters

This table reports determinants of a fund's decision to disclose a voluntary shareholder letter with qualitative disclosure using Logit regressions. Variables are defined in Appendix A. The z-statistics, in parentheses, are based on standard errors clustered by fund. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels, respectively.

	(1)	(2)	(3)	(4)
Dependent Variable		*ShareholderLetter*		
PastRisk	0.009	0.013	0.014*	0.020**
	(1.38)	(1.57)	(1.81)	(2.03)
PastAlpha	-0.003	-0.003	-0.000	-0.000
	(-1.18)	(-1.22)	(-0.14)	(-0.09)
PastFlow	-0.000	-0.000	0.000	0.000
	(-0.20)	(-0.22)	(0.52)	(0.46)
Log(TNA)	0.012	0.013	-0.021	-0.021
	(0.44)	(0.48)	(-0.63)	(-0.63)
Expense	-0.114	-0.115	-0.293	-0.308
	(-0.79)	(-0.80)	(-1.26)	(-1.31)
Turnover	0.001*	0.001*	0.001*	0.001*
	(1.86)	(1.89)	(1.89)	(1.91)
Log(Age)	0.257***	0.257***	0.194**	0.191**
	(4.04)	(4.03)	(2.55)	(2.52)
PastReturnGap	-0.028	-0.027	-0.037	-0.035
	(-0.92)	(-0.89)	(-1.09)	(-1.02)
Tenure			0.028***	0.028***
			(3.68)	(3.69)
12b1Fee			0.772*	0.815*
			(1.86)	(1.95)
MgmtFee			0.096	0.101
			(1.10)	(1.15)
Observations	24,048	24,048	17,755	17,755
Year FE	No	Yes	No	Yes
Pseudo R-squared	0.008	0.008	0.015	0.015