Do Investors Overreact to Managerial Tones?

WEI WANG*

November 11, 2022
[Latest Version]

ABSTRACT

This paper investigates whether investors overreact to managerial tones in the text of annual reports. Using a machine learning technique on topic analysis, I conduct a decomposition of managerial tones on multiple topics (signals), which are informative about different aspects of firms' value (fundamentals). Focusing on 10-Ks of SP500 constituents, I find the market reacts more to signals which are precise about future fundamentals, indicating that investors extract value-related information from tones while ignoring text uncorrelated with fundamentals. A simple structural model of trading shows that while naive investors overestimate excess returns by 2.75 basis points on average, investors' overreaction pushes up the filing period buy-and-hold excess return by only 0.03 basis points. Cross-sectional analysis shows that overreaction is most salient for high market valuation firms and medium-size ones. The results support the view that investors are not overreacting to managerial tones.

Key Words: Corporate Disclosure, Managerial Sentiment, Machine Learning, Textual Analysis
JEL classification: G12, G14, G41

*Wei Wang is with department of finance, Indiana University (Email: ww18@iu.edu). The views in this paper are the responsibility of the author, not the institutions he is affiliated with. I am indebted to Christian Heyerdahl-Larsen, Noah Stoffman, Jordan Martel, Charles Trzcinka, Wenyu Wang, Alessandro Previtero, Jun Yang, Preetesh Kantak, Kristoph Kleiner, and other participants in Kelley Finance Seminar and American Financial Association 2022 for supportive and helpful comments. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

Textual data in corporate disclosure get increasing attention from investors. Cohen, Malloy, and Nguyen (2020) shows that financial statements word number increases dramatically recently. The natural language processing technology augments investors' ability to price textual information (Cao, Jiang, Yang, and Zhang, 2020). However, in a survey of textual analysis, Loughran and McDonald (2016) note that: "Textual analysis, and more generally qualitative analysis, is most notably demarcated from quantitative analysis by its imprecision." The popularity and ambiguity of textual information make it important to study how investors understand and react to textual information in corporate disclosure.

This paper studies whether markets overreact to managerial tones. Efficient market hypothesis would say no. It argues that public information should be priced immediately. However, textual information needs more cognitive resources than quantitative information to digest and apply to investment decisions. Cohen et al. (2020) demonstrate that narrative structural changes are priced slowly and this ignorance is not exploited by arbitrageurs, indicating that investors "ignore" them due to the burden of text processing. Existing empirical results documents evidence on market reactions and return predictability of managerial tones but disagree on whether the mechanism is a rational story or not. With rational stories, market reaction is a reflection of information incorporation. Loughran and McDonald (2011) find negative words are correlated with market reactions with comprehensive controls, which is consistent with the theory that managerial tones are informative and investors price the information. However, Jiang, Lee, Martin, and Zhou (2019) document the negative predictability over future returns of managerial tones in corporate disclosure. Based on their statement, the negative predictability is a consequence of investors' naivety, "*...investors may simply follow managers' sentiment in financial disclosures, even though this sentiment may not represent the underlying fundamentals of the firm...*". The conflicting theories reflect the complex nature of this question on the empirical side. First, overreaction is a comparative statement. It needs a benchmark. It is empirically hard to build a "correct" benchmark. Second, investors may have private information, which is unobservable.

To answer this question, I propose an identification strategy relying on topic-level sentiments of corporate documents, which is accomplished by a modified version of textual analysis technique, Latent Dirichlet Allocation ("LDA"). My results are two-fold. First, I find the market reacts more to signals which are precise about future fundamentals, indicating that investors extract

value-related information from tones while ignoring text uncorrelated with fundamentals. Second, a simple structural model of trading shows that while uninformed investors overestimate excess returns by 2.75 basis points on average, investors' overreaction pushes up the filing period buy-and-hold excess return by only 0.03 basis points. Extra cross-sectional analysis shows that overreaction is most salient for high market valuation firms and medium-size ones. The results support the view that investors are not overreacting to managerial tones.

This paper makes two notable contributions to the literature. First, I propose a novel identification strategy to quantify whether the market overreacts. The strategy comes from a simple model. The model has investors with and without private information. The overreaction problem is transformed into the quantification of uninformed investors' relative magnitude, which is identifiable. I assume two types of investors. One has access to private information and trades on both private and public information (informed investors). The other has no access to private information and only trades on public information. I assume that prior knowledge is uninformative so uninformed investors will put all weights on public signals once they arrive, which looks like they take what managers say without thinking. The uninformed investor is a source of overreaction. To circumvent the direct measurement of private information, I focus on cross-sectional instead of time-series information . To be more specific, I look at market reactions to managerial tones on different aspects, such profitability, energy costs, etc. By calibrating the precision of topic-level managerial tones on fundamentals, I can compare the sensitivity of market reaction to different topics. If the sensitivity is higher for informative signals, it indicates that investors are mostly informed and use both private and public signals based on their relative precision. The heterogeneity in sensitivity can help back out the magnitude of informed investors.

Second, I develop a new topical sentiment extraction framework based on LDA and Loughran and McDonald dictionary to decompose managerial tones on multiple topics. This provides granular and comprehensive sentiment measures on documents. Previous endeavor on calculating managerial sentiments resides either on document-wise (Loughran and McDonald, 2011) and market-wide (Jiang et al., 2019) or some manually-selected dimensions (Hassan, Hollander, van Lent, and Tahoun, 2019, 2020b; Li, Shan, Tang, and Yao, 2020). My framework combines machine learning techniques for topic analysis with lexicon-based sentiment extractions. Granular sentiment measures retain more information because aggregation cancels out cross-group variation. On the

methodological side, it is an adaptation of LDA to business texts. LDA is an unsupervised algorithm and therefore has no need for manually-labelled training dataset. It labels each sentence so ensures that there is no omitted information or topic . Using 10-K files of SP500 firms from 1993 to 2018, topic analysis shows that text in annual reports contains signals on six aspects: stock compensation, sales, energy costs, debt burdens, profits, and others.

My paper contributes to the literature on sentiment analysis in economics and finance. Textual data in financial statements contain massive information. Early studies use lexicon-based methods to measure sentiments (Tetlock, Saar-Tsechansky, and Macskassy, 2008; Loughran and McDonald, 2011; Garcia, 2013). A shortcoming of this method is that semantic meaning of vocabularies changes with contexts. Lexicon-based methods do not perform well on this aspect. Recent studies leverage the development of machine learning literature and use either supervised or unsupervised learning algorithms to adapt to this concern (Ke, Kelly, and Xiu, 2019; Garcia, Hu, and Rohrer, 2020). Loughran and McDonald (2016) and Gentzkow, Kelly, and Taddy (2019) give a comprehensive review of related papers. Buehlmaier and Whited (2018) construct a naive Bayes network to measure firms' financial constraints. (Hassan et al., 2019, 2020b; Hassan, Hollander, van Lent, and Tahoun, 2020a) uses a computational linguistic tool to measure firm-level political risk, Brexit risk, and Covid-19 risk. Li et al. (2020) use the same tool to extract firms' climate risk. While these measures are of great academic value, validation on measure accuracy should not be ignored for application.

My paper also relates to literature on LDA variants and their applications in finance. Lopez-Lira (2019) and Grundy and Petry (2020) restore firms' risk factors from the corresponding sections in 10-K documents using LDA. Fedyk and Hodson (2019) apply LDA to analyze resumes contents of millions of employees from US public firms and categorize the skill sets into different groups. Filatov (2020) uses FOMC minutes transcripts to extract topical themes. Bybee, Kelly, Manela, and Xiu (2020) quantify states of the economy by extracting news attentions in Wall Street Journal articles. Brown, Gredil, and Kantak (2019) analyze meeting scripts for conference between asset allocators and hedge fund managers. Liu, Sheng, and Wang (2020) apply this technique to ICO Whitepaper and propose an index measuring the fundamental value of cryptocurrencies. Huang, Lehavy, Zang, and Zheng (2018) use LDA to analyze the information asymmetry in analysts' reporting. The framework in this paper is slightly different from the original version of LDA proposed in Blei,

Ng, and Jordan (2003), which contributes to machine learning literature of aspect-based sentiment analysis. Moghaddam and Ester (2012) propose Joint Sentiment/Topic Model to extract topic and document-level sentiment at the same time. Based on JST, Lin, He, Everson, and Ruger (2011) propose reverse JST and extract topic-level sentiment. My framework shares the same spirits with these two models but more adapted for long documents.

The remainder of this paper is organized as follows. Section I presents a static micro structure model and develops the econometric setting. Section II discusses the estimation results. Section III concludes.

I. The Model

This section constructs a multi-period model. The general goal is to develop a framework to define "overreaction" and rationalize price reversion from the Bayesian updating theory. In the model I introduce informed and uninformed investors, who have differentiated access to private information. Understanding the source of price reversion helps develop an econometric setting to quantify the relative role of two types of investors.

A. Model Setup

Consider a model with one risky asset. The time frame is between 0 and $2T$. The asset is liquidated at the end. The liquidation value μ is not observable. The prior belief on μ is $N(\mu_0, \tau_0^{-1})$. To highlight the effect of external signals, I assume that $\tau_0 \to 0$ (i.e., very close to zero). I also assume there are two types of investors in the stock market: informed and uninformed, with $1-\chi$ and χ as the proportion to the entire population respectively. The only difference is that informed investors have access to private information while uninformed investors do not have. Under the Bayesian updating, the uninformed investors behave like replacing the beliefs with public signals when the prior is purely uninformative.

The private signal in period t has the following form: $s_t = \mu + e_{s,t}$ where $e_{s,t}$ is normally distributed with zero mean and τ_s precision (the inverse of variance). At time $T^* \in (T, T+1)$, a public signal becomes observable to all investors, denoted as F_T. $F_T = \mu + e_{f,T^*}$ with e_{f,T^*} normally distributed with the mean μ and precision τ_f. Public signals are observable to both informed and

uninformed investors.

B. Information Structure and Belief Updating

Before the arrival of public signals, informed investors update their beliefs based on private signals while uninformed investors do not update their beliefs. The standard Kalman filter theory demonstrates that $\mu|\{s_i\}_{i=1}^t$ is normally distributed with mean m_t and precision τ_t. It can be easily verified that

$$m_t = \begin{cases} \frac{1}{t}\sum_{i=1}^{T} s_i & \text{if } t \leq T \\ \frac{\tau_s}{T\tau_s+\tau_f}\sum_{i=1}^{T} s_i + \frac{\tau_f}{T\tau_s+\tau_f}F_T & \text{if } t = T^* \\ \frac{\tau_s}{t\tau_s+\tau_f}\sum_{i=1}^{t} s_i + \frac{\tau_f}{t\tau_s+\tau_f}F_T & \text{if } t \geq T+1 \end{cases} \tag{1}$$

$$\tau_t = \begin{cases} t\tau_s & \text{if } t \leq T \\ T\tau_s + \tau_f & \text{if } t = T^* \\ t\tau_s + \tau_f & \text{if } t \geq T+1 \end{cases} \tag{2}$$

The uninformed investors

$$\tilde{m}_t = \begin{cases} \mu_0 & \text{if } t \leq T \\ F_T & \text{if } t = T^* \\ F_T & \text{if } t \geq T+1 \end{cases} \tag{3}$$

$$\tilde{\tau}_t = \begin{cases} \tau_0 & \text{if } t \leq T \\ \tau_f & \text{if } t = T^* \\ \tau_f & \text{if } t \geq T+1 \end{cases} \tag{4}$$

The beliefs of informed investors are updated on private and public information while those of uninformed investors only change on public information. At the arrival of the public signal, beliefs of informed and uninformed investors change differently.Informed investors weight the current posterior beliefs and new public signal by their respective precision. Uninformed investors put high weights on public signals due to the uninformativeness of prior beliefs and behave like replacing the

existing beliefs with the signal. This behavior feature is consistent with the description of naive investors in Jiang et al. (2019).[1]

C. *Portfolio Choices and Equilibrium*

Investors are risk-neutral and maximize expected net trading profits each period. Trading profits include the gross trading profits, which is the number shares bought times the gap between the liquidation value and the stock price, and trading costs. The objective function of all investors has the same form:

$$E_t[x_t(\mu - p)] - \frac{\eta}{2}x_t^2 \tag{5}$$

where x_t is the trading volume at time t and η is the trading costs.The quadratic term describes the increasing marginal costs from trading, reflecting price impacts. The optimal instantaneous trading strategy is solved as $x_t = \frac{1}{\eta}(E_t[\mu] - p)$. For convenience, denote m_t for informed investors' $E_t[\mu]$ and $\tilde{m}_t$ for uninformed investors.

Suppose the supply of asset u_t is a normally distributed random variable with mean 0 and variance τ_u^{-1}. In equilibrium, the asset market clears:

$$(1-\chi)\frac{1}{\eta}(m_t - p_t) + \chi\frac{1}{\eta}(\tilde{m}_t - p_t) = u_t \tag{6}$$

The equilibrium asset price is $p_t = (1-\chi)m_t + \chi\tilde{m}_t - \eta u_t$.

The next proposition states the price behaviors around the arrival of public signals and in the post-announcement period (i.e., after the arrival of public signals).

PROPOSITION 1: *Around the arrival of public signals, the price behaviors in the short run and long run have the following features. When* $\chi < 1$.

- *Market reaction:* $\frac{\partial}{\partial\chi}\left(P_{T^*} - P_T\right) = \frac{T\tau_s}{T\tau_s+\tau_f} > 0$
- *Price reversion: when* $t > T$
  1. $\frac{\partial}{\partial F_T}\left(P_{t+\Delta} - P_t\right) = \left[\frac{1}{(t+\Delta)\tau_s+\tau_f} - \frac{1}{t\tau_s+\tau_f}\right]\tau_f < 0$

[1] Jiang et al. (2019) claim that "*...investors may simply follow managers' sentiment in financial disclosures, even though this sentiment may not represent the underlying fundamentals of the firm...*".

2. $\frac{\partial^2}{\partial\Delta\partial F_T}(P_{t+\Delta} - P_t) = -\frac{\tau_s\tau_f}{[(t+\Delta)\tau_s+\tau_f]^2} < 0$

3. $Corr(P_{t+\Delta} - P_t, F_T) < 0$

The first item in Proposition 1 highlights why χ is an important parameter to estimate. The reason is that it is a measure of overreaction. The proposition shows that the market reaction to the public signal is more salient when the uninformed investor accounts for larger portions. This is because uninformed investors update the posterior beliefs by putting all weights on public signals.

The second item demonstrates that informed investors' behavior is the only source of price reversion. Notice that The price reversion does not exist when $\chi = 1$. In the long run, informed investors finally learn the true value and correct the noises in previous signals as τ_t goes to infinity as $t \to \infty$. Any noises in public signals will be corrected. Point 3, $Corr(P_{t+\Delta} - P_t, F_T) < 0$, reflects this process of belief updating using new information. Even when $\chi = 0$, price reverts after the public signal arrives. Private information arrives each period and dilutes the weight of public information in the posterior beliefs as shown in Point 1 and 2.

The pattern of price reversion is considered as an evidence against efficient market hypothesis. The reversion is treated as a delayed incorporation of information into price which is available at the public signal arrival. In my framework, the public information is one-time but private information flows constantly. When public signals arrive, stock prices react immediately. After that, markets continue updating the price using the private information constantly. Notice that price reversion exists even when all investors are rational. This is because the price reversion is a consequence of learning and correcting.

To summarize the results and implications from the model, proposition 1 shows that as the uninformed investor accounts for higher proportion, the market reaction to public signal is more salient, the price reversion is less salient. The public signal can predict future returns. Two points are worth noting. First, those results are not conflicting evidence against EMH. Semi-strong version of EMH claims that prices incorporate all public available information, which still holds in the model. The price reversion is due to the correlation between public signals and private signals arriving later. If there is no private information after the arrival of public signals, there would be no price reversion. Second, the predictability exists because I assume that the market clears each period in a static way. My argument on this assumption is the pervasive limits of arbitrage (Shleifer

and Vishny, 1997). When arbitrage is costly, investors cannot fully eliminate the predictability by taking advantage of it.

Section II develops an econometric strategy to estimate the magnitude of χ by focusing on T^* in a setting where μ is dependent on multiple determinants. I explain why not relying on time-series information and why extending the model to multi-factor version.

II. Empirical Tests

The magnitude of χ is the parameter to be estimated but the quantification is challenging for two reasons. First, many of the variables in the model are not measurable. For example, private information and investors' beliefs are not observable to econometricians. The concepts "signals" and "liquidation values" are hard to find counterparts empirically. The second reason is the empirical model misspecification. For example, I assume the precision of private signals is constant before and after the arrival of public signals, which might be time-varying in reality.

To solve the first challenge, I extract signals from textual data in corporate disclosures such quarterly reports or annual reports. Textual data is high-dimensional, containing various topics. Each topic provides a unique snapshot of firms' fundamental. And by looking at public signals, I don't need to measure private signals explicitly. For the second challenge, I propose an identification strategy using only the cross-sectional information at the arrival of public signals (i.e., when firms publish documents). Focusing on cross-sectional information circumvents the potential time-varying nature of signals. The market reaction to public signals is also the most important setting where market efficiency is tested.

A. Identification Strategy

Focusing on the moment at T^*, I extend the basic model. Different from the original setting, now the asset value, μ, is determined by N i.i.d random variables, i.e., $\mu = \sum_{i=1}^{N} w_i \delta_i$. δ_is are determinants of the asset liquidation value, such as profitability, default risks, etc. [2]. δ_is are not observable due to the information asymmetry between insiders and outsiders so investors rely on private signals and corporate disclosures to update their beliefs on δ_i. The prior belief of δ_i is

[2]I use "determinants" instead "factor" to distinguish from the terminology widely used in asset pricing literature

$N(\mu_i, \sigma_i^2)$.

Informed investors receive N private signals. Each signal conveys information on one determinant: $s_i = a_i + b_i\delta_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_{\varepsilon,i}^2)$. ε_i is the noise term independent of all other variables, including factors and other noises. The linear form of signals on fundamentals accounts for the fact that signals and fundamentals are not on the same scale. a_i and b_i can be seen as scaling coefficients.

Using the property of Gaussian distribution, we know that $f_i|s_i \sim N(\mu_i^*, \Sigma^*)$, where $\mu_i^* = \frac{1}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2}\left[\frac{b_i^2}{\sigma_{\varepsilon,i}^2}\frac{s_i-a_i}{b_i} + \frac{1}{\sigma_i^2}\mu_i\right]$ and $\Sigma^* = \frac{1}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2}$. Therefore, informed investors will update their beliefs by the following rule:

$$\sum_{i=1}^{N} w_i \frac{1}{1/\sigma_i^2 + b_i^2/\sigma_{\varepsilon,i}^2}\left[\frac{b_i^2}{\sigma_{\varepsilon,i}^2}\frac{s_i - a_i}{b_i} + \frac{1}{\sigma_i^2}\mu_i\right]. \tag{7}$$

Similar to Section I, the uninformed investors will update the belief on μ by putting all weights on the public signals: $\sum_{i=1}^{N} w_i \frac{s_i-a_i}{b_i}$. The asset price in equilibrium is:

$$p = \chi\sum_{i=1}^{N} w_i\frac{s_i - a_i}{b_i} + (1-\chi)\sum_{i=1}^{N} w_i \frac{1}{1/\sigma_i^2 + b_i^2/\sigma_{\varepsilon,i}^2}\left[\frac{b_i^2}{\sigma_{\varepsilon,i}^2}\frac{s_i - a_i}{b_i} + \frac{1}{\sigma_i^2}\mu_i\right] - \eta * u. \tag{8}$$

The first and second term in equation (8) correspond to the expectation of asset value for uninformed and informed investors. If we reorganize the term,

$$p = \chi\sum_{i=1}^{N} w_i \frac{1}{1/\sigma_i^2 + b_i^2/\sigma_{\varepsilon,i}^2}\frac{1}{\sigma_i^2}\left(\frac{s_i - a_i}{b_i} - \mu_i\right) + \sum_{i=1}^{N} w_i \frac{1}{1/\sigma_i^2 + b_i^2/\sigma_{\varepsilon,i}^2}\left[\frac{b_i^2}{\sigma_{\varepsilon,i}^2}\frac{s_i - a_i}{b_i} + \frac{1}{\sigma_i^2}\mu_i\right] - \eta * u. \tag{9}$$

In the empirical part, I employ maximum likelihood estimation and estimate the following model

$$r_t = \alpha + \sum_{i=1}^{N} \tilde{w}_i[\tilde{s}_i + k\tilde{s}_i^*] + X_t\Phi + \varepsilon_t \tag{10}$$

where $\tilde{w}_i = \chi w_i$, $k = \frac{1-\chi}{\chi}$, $\tilde{s}_i = \frac{s_i-a_i}{b_i}$, $\tilde{s}_i^* = \frac{1}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2}\frac{b_i^2}{\sigma_{\varepsilon,i}^2}\frac{s_i-a_i}{b_i} + \frac{1}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2}\frac{1}{\sigma_i^2}\mu_i$. X_t include other covariates that can predict returns. This term captures the hard information that investors can get

to form their prior. The goal is to estimate α, $\tilde{w}_i$, k, Φ.

The identification of χ is highly reliant on the heterogeneity assumption of relative precision $((\sigma_i b_i)/\sigma_{\varepsilon,i})$ across signals. Suppose $(\sigma_i b_i)/\sigma_{\varepsilon,i}$ is a constant c for all i. Equation 10 will degenerate to $r_t = \alpha + \sum_{i=1}^{N} \tilde{w}_i(1 + \frac{k}{1+1/c^2})\tilde{s}_i + X_t\Phi + \varepsilon_t$. $\tilde{w}_i$ and k cannot be identified separately. This assumption is not very strong and therefore quite easy to be satisfied. It requires the relative magnitude $\frac{b_i}{\sigma_{\varepsilon,i}}$ and $\frac{1}{\sigma_i}$ remains heterogenous. It means that unless they comove across signals perfectly this assumption can hold.

What is left is the construction of signals, $\tilde{s}_i$ and $\tilde{s}_i^*$. Section II.B will introduce the variant of LDA and how to extract signals using this algorithm.

B. Textual Analysis Framework

Even though regulations have requirements and formats on disclosure materials, managers have discretion on the vocabulary choices, sentence structures, etc. And the topics are disperse physically and diverse across firms. The same document may be only informative on one or several topics. That makes manual labeling extremely laborious. So I leverage one technique developed in computational linguistics research called Latent Dirichlet Allocation and accomplish two goals: (1) classify texts into different topics; (2) quantify the sentiment on that topic.

LDA is first proposed in Blei et al. (2003) and has been widely used in topic analysis. It is an unsupervised learning and requires no training. The idea is to identify tokens that appear together frequently and classify them into one group. I refer details to the original paper. But that version cannot directly apply to our question because we need both topics and sentiment measures. Therefore I make some modifications on the base version. I first group the entire token set into two categories: sentiment words, a subset of vocabularies or phrases that are strong in lexicon polarity, and neutral words, which are relatively neutral in polarity but speak more on material themes. I define tokens included in the negative v.s. positive word lists first developed in Loughran and McDonald (2011) as sentiment words. Typical examples are "undetermined" or "distortion". Tokens not in the list are defined to be neutral words. The variant's idea is that I apply LDA to neutral words and let sentiment words variate across topics. Once a sentence is classified as talking a certain topic, the sentiment words in that sentence tell about the sentiment.

My framework uses the bag-of-words model. Following the traditional notations, each topic is

a distribution over neutral words. A certain topic may put high probability mass on some words. Documents are collection of sentences, which is a collection of tokens. A slightly difference here is that I assume tokens in one sentence share the same topic so that the collection of sentences give a distribution of topics in a certain document. This assumption is to increase the accuracy when the model associates sentiments with topics. Without this assumption, the sentiment words in the first sentence could be associated with the topic in the last sentence, which is not precise. This assumption incorporates the prior that physical approximation is informative. I assume that sentiment words' distributions are heterogeneous across document and topic. The sentiment words in that sentence can measure the managerial tone on that topic. The distributions on sentiment words are informative of managerial tones on that topic, based on which I construct topic sentiment measures. This variant is close to Lin and He (2009); Lin et al. (2011) in terms of their ideas adding extra layers to capture more textual variations.

Section II.B.1 will describe the notations and generative process of our model. Section II.B.2 presents the procedure of Gibbs Sampling for model estimation.

B.1. Model Setup

For notational clarity, suppose each document d composes of U_d sentences and sentence u composes of $N_{d,u}$ tokens. Token selection process is governed by what themes authors want to discuss (topic) and what attitude they hold against it (sentiment). A topic is a probability distribution over neutral words, $\mathcal{N}$. In each document, distributions over sentiment words, $\mathcal{S}$, are different across topics. The statistical representation of topic or sentiment as a probability distribution indicates that different topic or sentiment will have "preference" over a subset of vocabularies.

Following the conventions in LDA, for document d topics are selected following a multinomial distribution (topic mixture) with parameters $\theta_d = \{\theta_{d,t} | t \in \{1, \ldots, T\}\}$. These parameters sum up to one and relative magnitudes indicate how much attention authors pay on each topic. Document d's sentiment of topic t is denoted by $\varphi_{d,t} = \{\varphi_{d,t,v} | v \in \mathcal{S}, \varphi_{d,t,v} \geq 0, \sum \varphi_{d,t,v} = 1\}$. Unlike topics, we do not explicitly model sentiment as a parameter. Instead, we do this indirectly by modeling sentiment as a distribution on sentiment words, i.e, $\varphi_{d,t}$. This will provide more flexibility for sentiment index calculation. We also assume that each sentence discusses only one topic, $z_{d,u}$. The generative process of this model is summarized in Definition 1.

DEFINITION 1: *The generative process is as follows:*

1. *For each document d, choose a topic mixture* $\theta_d \sim Dir(\alpha)$ *and the probability of using sentiment word for each topic* $\zeta_{d,z} \sim Beta(\nu, \nu)$
2. *Under document d, for each topic t, choose a sentiment mixture* $\varphi_{d,t} \sim Dir(\beta)$
3. *For the sentence u in document d*
   (a) *choose a topic* $z_{d,u} \sim Multinomial(\theta_d)$
   (b) *For the token n, where* $n = 1, \ldots, N_{d,u}$
      i. *choose whether* $w_{d,u,n}$ *belongs to* $\mathcal{S}$ *and* $\mathcal{N}$ *following Bernoulli distribution,* $Bern(\zeta_{d,z_{d,u}})$.
      ii. *If it is a sentiment word, choose a word* $w_{d,u,n} \sim Multinomial(\varphi_{d,z_{d,u}})$. *Otherwise, choose a word* $w_{d,u,n} \sim Multinomial(\phi_{z_{d,u}})$

[Place Figure 1 about here]

Before diving into estimation, a few discussions on this variant of LDA. Why do we need to use LDA instead of other techniques? LDA is an unsupervised learning algorithm. Unsupervised learning algorithm requires no human labelling. So we can input all the narratives into the model and let computers label them. The output is based on all the texts in corporate disclosure. This comprehensiveness can provide a full decomposotion of overall managerial tones. Hassan et al. (2020a) use a different technique to extract sentiment measures on different dimensions. But their algorithm needs manual checkings to select which dimensions to extract. The framework in this paper fits the question better.

Next, why do we need to propose a variant of LDA instead of using LDA directly on the sentence level? Huang et al. (2018) adopt a simple alternative way. They first split documents into sentences and run LDA taking each sentence as a document. They assign each sentence to the most likely topic. Then they calculate the sentiment of that sentence and aggregate into document-topic level. This method is easy to implement. But there might be some concerns for applicability in other cases. The most notation concern is that the output of LDA on sentence topics is a distribution over topics. So it could be that the distribution is quite even, indicating LDA cannot classify topics with a high confidence level. Another concern is that while LDA is a Bayesian framework, their way is not Bayesian. This mixed strategy of application is flexible but may also lead to confusion

of statistical properties. The framework in this paper shares similar ideas, including taking one sentence as the unit and aggregating into document-topic level, but adopt a complete bayesian framework.

There are two features of this model: first, it assumes that each sentence has one topic; second, it groups only on neutral words, based on which topics are classified. These two assumptions can provide the following benefits: (1) greater convergence speed; (2) more accurate classification. Gains on convergence speeds are from incorporating the prior that proximity by distance is indicative of topic similarity. Tokens appearing together are more likely to be under one topic. This reduces the requirement of input document sizes. Furthermore, grouping only on neutral words can use the most relevant information. Sentiment words are more indicative of polarity information instead of theme-related information. Exclusion of them can denoise the training sample. The estimation method is the same as LDA, which is discussed in Section II.B.2.

B.2. Model Estimation

The model is estimated using collapsed Gibbs sampling, i.e., first sampling topic of each sentence Z, then estimating the parameters given the sampled Z. Steyvers and Griffiths (2007) show that the parameters for estimation can be recovered if latent variable Z is observable for LDA.

The variables of interests include $\zeta_{d,t}$, ϕ_t, θ_d, and $\varphi_{d,t}$ for $d = 1, \cdots, D$ and $t = 1, \cdots, T$. The goal is to estimate the posterior distribution of $\zeta_d|W_d$, $\phi_t|W$, $\theta_d|W_d$, and $\varphi_{d,t}|W_d$. I calculate the matrix norm of $K_n'(t) - K_n(t)$, which is the discrepancy between two consecutive updated matrix $K_n(t)$. This measure will be used to check the convergence of Gibbs sampling. Appendix A presents the detailed algorithm. Once Z is sampled, proposition 2 gives the posterior distribution of parameters. Corollary 3b indicates how parameters can be estimated.

PROPOSITION 2: *Given the latent variable Z is observed: (1) $\zeta_{d,t}|(z_d, W) \sim Dir(N^{\mathcal{S}}_{d,t}+\nu, N^{\mathcal{N}}_{d,t}+1)$; (2) $\theta_d|(z_d, W) \sim Dir(C(d)+\alpha)$; (3) $\varphi_{d,t}|W, z_d \sim Dir(K_s(d,t)+\beta)$; (4) $\phi_t|W, Z \sim Dir(K_n(t)+\gamma)$ where $C(d)+\alpha = (c(d,1)+\alpha, \cdots, c(d,T)+\alpha)$, $K_s(d,t)+\beta = (k_s(d,t,1)+\beta, \cdots, k_s(d,t,V_s)+\beta)$, $K_n(t)+\gamma = (k_n(t,1)+\gamma, \cdots, k_n(t,V_n)+\gamma)$. $c(d,t)$ is the number of sentences in document d assigned to topic t. $k_s(d,t,v)$ is the number of times sentiment word v is assigned to topic t and sentiment s. $k_n(t,v)$ is the number of times sentiment-neutral word v is assigned to topic t.*

COROLLARY 1 (Parameter Estimate): *Given the latent variable Z is observed, θ, π, ϕ, and φ can be estimated by:* $\hat{\zeta}_{d,t} = \frac{N^{\mathcal{S}}_{d,t}+\nu}{N_{d,t}+2\nu}$, $\hat{\theta}_{d,t} = \frac{c(d,t)+\alpha}{\alpha T+\sum c(d,t)}$, $\hat{\varphi}_{t,s,v} = \frac{k_s(d,t,v)+\beta}{\sum k_s(d,t,v)+\beta*V_s}$, *and* $\hat{\phi}_t = \frac{k_n(t,v)+\gamma}{\sum k_n(t,v)+\gamma*V_n}$.

Corollary 3b shows how the estimates are constructed from the sampled topic. Suppose prior distribution plays no role, the estimate is just the sample-moment analogy of parameters. I will set the hyperparameters close to zero to reduce the effect of prior information.

C. Data and Samples

I restrict the sample to 10-K and 10-K405 documents provided by Loughran and McDonald Stage-One 10-X Parse files updated to 2018. The Stage One Parse excludes markup tags, ASCII-encoded graphics, and tables. Table I shows how the original sample of 10-Ks is impacted by the data filters and data requirements. I follow the filters imposed in Loughran and McDonald (2011) but focus only on SP500 constituents.

[Place Table I about here]

Before feeding the model with corpus, I perform a set of standard cleaning procedures. The goal is to only keep informative words. First, I eliminate all number characters, punctuation, and anything that are not alphanumeric characters. Then I remove words with length less than 3 and those in all stopword lists provided by Loughran and McDonald's webpage. I include a handful of terms into this stopword list. I exclude those words that only appear in one document. I only keep words that appear at least in two sentences. Since the LM dictionaries are unstemmed, I will present my results using unstemmed words. Details are provided in Appendix C.

I use the positive and negative words lists developed in Loughran and McDonald (2011) for sentiment words. The remaining words are all classified as neutral words. The final sample contains 12,028 documents from 976 unique firms (identified by COMPUSTAT item *gvkey*). There are 2,340 unique sentiment words and 112,961 neutral words in sample. It contains 17,051,664 sentences with 10,571,093 sentiment words and 225,382,330 neutral words. Table II presents the summary statistics. *% Positive* and *% Negative* are compared to Table II in Loughran and McDonald (2011).

[Place Table II about here]

D. *Topic Analysis*

The LDA model needs manual setting of topic number. The following results are based on the output when this parameter is set to six. The selection of this parameter value depends on corpus and there is no standard procedure. The main consideration is that the output is human-interpretable enough. The burning periods are set to be 150 when the Gibbs sampling demonstrates stable outputs. Burning periods are iterations for models to reach stable distribution. The algorithm will sample another 100 times and take the data every 5 times for estimation. This is to avoid sequential correlation. I set the hyperparameters to be close to zero so that the effect of prior information can be minimized. $\alpha = \beta = \gamma = \nu = 0.0001$. Appendix B presents tokens with 30 highest probabilities and their histogram in Figure 5. Notice that some words like *COMPANY* are very common in six topics mainly because the model assumes that one sentence has one topic. *COMPANY* often shows as subjects or objects in many topics. When labelling the topic, words that are relative unique give more clues.

The first topic includes many terminologies and is very inclusive. For example, *CONSOLIDATED* appears in phrases like *consolidated balance sheets*, *consolidated balance sheets*, etc. *INTERNAL CONTROL* often appears in the phrase *Opinion on Internal Control*. Accounting firms will deliver their opinions on auditing firms' annual reports, which is summarized in Section *Report of Independent Registered Public Accounting Firm* [3]. *EXCHANGE ACT* often appears in phrases like *SECURITIES EXCHANGE ACT* when regulations are referred for disclosures. I give it a label "General". It proxies for overall disclosed information without specific focuses.

The second topic is related to compensation, as *COMPENSATION* or *AWARDS* appear in the top list. The words like *SHARES*, *OPTIONS*, *RESTRICTED* indicate that this topic focuses on equity-based compensation. *EMPLOYEES* and *DIRECTORS* show that it speaks to not only directors or executives but also rank-and-file employees. So I label this topic as "Compensation", which proxies for corporate equity-based compensation.

The next topic centers on product & market related issues. *PRODUCT*, *SERVICES*, *CUSTOMERS*, *MARKET*, and *OPERATIONS* imply that this topic is about businesses' operating issues. *GROWTH*, *INCREASED*, *REVENUE* refer to the performance of business operation. So I

[3] See APPLE. Inc Sep, 2020, 10-K for an example

will label this topic "Sales", proxying for product market performance.

The following topic hinges on one dimension of costs, which is getting more attention recently. *NATURAL GAS*, *OIL*, *ELECTRIC*, and *FUEL* are terms of energy types. *PRODUCT COSTS*, *PRICE*, *INCREASE*, and *FUTURE* reflect that companies report that increase in energy prices can add to current and future production costs. Based on this reasoning, I label this topic "Energy", proxying for energy-induced product costs.

The fifth topic discusses debt financing costs. The topic explicitly mentions *DEBT* and *LOANS* as well as debt-related expenditure like *INTEREST RATE* and *DUE*. This topic should be labelled as "Debts", proxying for companies' debt burden.

The last topic related to the statement of income for its extensively mentioning terminologies like *INCOME*, *TAX*, *NET ASSETS*, *EXPENSE*, *COSTS*, etc. For clarify, I label this topic as "Profits", proxying for net margins.

[Place Figure 2 about here]

In Table II we can see that topic "Sales", "Debts", and "Profits" account for about sixty percent. Figure 2 demonstrates the dynamics of topic weights, drawing median, 5 percentile, and 95 percentile by year. While most topics are losing weight, weights on topic "Debt" and "Profits" are increasing.

E. Topic-level Sentiment

This section discusses how to calculate the sentiment of each topic on the document level. As mentioned before, the model summarizes the sentiment information in sentiment words distribution, $\hat{\varphi}_{d,t}$. Loughran and McDonald (2011) use the TF-IDF weighted negative word frequency as a sentiment index. Jiang et al. (2019) use the difference between the number of positive words and the number of negative words scaled by the total word count. These methods account for at least two factors: (1) the relative frequency of positive and negative words; (2) the proportion of sentiment words to total words.

The model in Section II.B can generate two relative paramters: (1) sentiment word proportion $\hat{\zeta}_{d,t}$; (2) sentiment word distribution $\hat{\varphi}_{d,t}$. Suppose $\eta = [s_1, \cdots, s_{V_s}]^T$ denote the sentiment score for each sentiment word, I assign $+1$ for positive words and -1 for negative words. The sentiment score

for document d and topic t, which is the proxy for signal s_i, is defined as: $\hat{s}_{d,t} = \hat{\theta}_{d,t}\hat{\zeta}_{d,t} * \eta^T * (\hat{\varphi}_{d,t} - I)$. There are three elements in this definition. First, $\eta^T * (\hat{\varphi}_{d,t} - I)$ is the average of token score weighted by their probability distribution. The subtraction of constant matrix is for normalization because Loughran and McDonald dictionary has longer list for negative words (2,355) than positive words (354). $\hat{\zeta}_{d,t}$ is the probability of sentiment word appearance. $\hat{\theta}_{d,t}$ is the weight of topic. If the document puts little weight on a topic or contains few sentiment words, sentiment information is too little to be accurate. These two elements account for the role of sample size on measurement accuracy.

The identification strategy is reliant on the heterogeneity of signals' precision in terms of containing future information. To calibrate signal precision, for each firm I run the time-series regression $\hat{s}_{i,t} = a_i + b_i f_{i,t+1} + \varepsilon_{it}$ for each topic i. $f_{i,t+1}$ is a measure based on future accounting information. I use forwarded factor value as the independent variable to capture the idea that the managerial tones are functions of future fundamental. The time-series regression will yield estimates of $\hat{a}_i$, $\hat{b}_i$, and $\hat{\sigma}^2_{\varepsilon,i}$. The fundamental uncertainty is proxied by $\hat{\sigma}^2_i = var(f_{i,t})$.

I choose the ratio of stock compensation expense to nominal book value for topic "Compensation". Equity-based compensation is to provide incentives for both managers and other employees (**?**). The level of equity-based compensation can measure the incentive provided to firms' employees. Topic "Sales" is about product market situation. I use sales growth as a proxy. I also use return on equity for Topic "profitability". The topic "Debt" mentions corporate debt burden, including interest payment, liability, etc. So following **?**, I calculate the average cost of debt, which is defined as Item XINT/(Item DLTT + Item DLC), to measure how costly for firms to assume debt. Topic "General" and "Energy" are challenging to get proxies, mainly because there are no specific items on financial statements. Since energy prices will affect firms' production costs, I project the total operating cost, Compustat Item XOPR, on the global price of energy index using linear regressions and then scale it with book value of assets, COMPUSTAT item AT. The time series of global price of energy index is available in Federal Reserve's database [4]. Topic "General" works as a kitchen sink, absorbing all topics that are not classified into five topics. To avoid confusion, I only add the sentiment score of topic "General" as a control variable.

[4] https://fred.stlouisfed.org/series/PNRGINDEXM

[Place Figure 3 about here]

Figure 3 plots the dynamics of sentiment scores on each topic and five proxy variables. Unsupervised learning algorithm needs manual interpretation. Cross-checking the trend of sentiments and proxy variables can validate the interpretations of topics. For example, Topic "Compensation", "Sales", Profits" and their proxies are positively correlated, with the correlation coefficents 0.2125, 0.6405, 0.3741. On the other hand, topic "Energy" and "Debts" are negative correlated with other proxies, with the correlation coefficients -0.0279 and -0.3047. The sign of correlation coefficients make sense because when "Energy Cost" and "Cost of Debt" are higher, firms will suffer higher burden of energy expenses and debt financing costs, which is not in favor of firms, indicating that they should be negatively correlated.

After firm-level calibration, we are able to get the weighting term, $\frac{b_i^2/\sigma_{\varepsilon,i}^2}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2}$, which proxies for how much weight a sophisticated investor would put on the signal. Table III Panel A show that Topic "Sales" gets the highest weight, about 6 percent. We want to know how much difference this weighting scheme would create between $\tilde{s}_i$ and $\tilde{s}_i^*$. Table III Panel B presents the correlation matrix. The correlation between $\tilde{s}_i$ and $\tilde{s}_i^*$ is below 10 percent. This means that variations of $\tilde{s}_i$ and $\tilde{s}_i^*$ are indicating quite different "information". Trading against one of them yields different outcomes and facilitates our identification. The assumption for the identification strategy is satisfied.

[Place Table III about here]

F. Estimation Analysis

Table IV replicates the main results in Loughran and McDonald (2011). Their results show that one percent increase in negative word proportion leads to 19.5 basis points reduction in filing period excess return. The coefficient for this sample is 15.6 basis points. If we replace the simple negative word proportion with sentiment scores of our six topics, we find that Topic "Sales" sentiment negatively correlates with excess return while "General" sentiment shows positive correlation as shown in Table V. Notice that these sentiments contain both fundamental information and noise. We want to see if signals are denoised whether the predictability still remains.

[Place Table IV about here]

[Place Table V about here]

Table VI replaces the independent variables with descaled signals $\tilde{s}_i$ and corrected and descaled signals $\tilde{s}_i^*$. In column (1), the predictability of all coefficients are very weak while after being corrected, this predictability greatly improves in column (3). $\tilde{s}_i^*$ for "Sales" is negatively correlated with excess return while $\tilde{s}_i^*$ for "Profits" positively correlates with excess return. Their predictability remains after controlling $\tilde{s}_i$ in column (5).

Notice that the only difference between $\tilde{s}_i$ and $\tilde{s}_i^*$ is that the latter is weighted by relative precision. The results that $\tilde{s}_i^*$ has higher predictability are consistent with the hypothesis that markets make sophisticated belief updating after receiving information from new disclosures.

[Place Table VI about here]

Now I will use the structural model developed in Section I to estimate the parameters in equation (10) $\{\alpha, \tilde{w}_1, \cdots, \tilde{w}_5, \Phi, k, \sigma^2\}$ by maximization of the following log-likelihood function:

$$l = \sum_{j,t} \left[-\frac{1}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} (r_t^j - \alpha - \sum_{i=1}^{N} (\tilde{w}_i \tilde{s}_{it}^j + \tilde{w}_i \tilde{s}_{it}^{*j}) - X_t^j \Phi)^2 \right] \tag{11}$$

subject to the constraint $k \geq 0$. The homoscedasiticity is for computing speed. I report 90% empirical bootstrapping confidence interval by bootstrap sampling 1,000 times for each estimate parameter. Bootstrap sampling suits this estimation because there is a variable transformation. Bootstrapping can avoid the approximation errors using Delta method.

[Place Table VII about here]

The results are summarized in Table VII. The confidence interval for χ is very small, indicating that the market is overall sophisticated, which is consistent with results in Table VI. In order to get a sense of economic magnitude, I conduct a counter-factual experiment. We can get the counterfactual filing-period excess return when the market is purely rational, i.e., $\chi = 0$. Then we can quantify how investors' naivity affects market reactions, $E[r_t - r_{t,\chi=0}]$. Finally, we can estimate how much naive investors overestimate returns, $E\left[\sum_{i=1}^{N} w_i \frac{1}{1/\sigma_i^2 + b_i^2/\sigma_{\varepsilon,i}^2} \frac{1}{\sigma_i^2} \left(\frac{s_i - a_i}{b_i} - \mu_i \right) \right]$.

Table VIII shows the results. If we look at the entire sample, naive investors have higher expectation of returns. Compared to sophisticated investors, their perception is 2.75 basis point higher. However, the proportion of naive investors is very low, about 1 percent. Therefore, overreaction from naive investor only pushes up the excess return by 0.029 basis point. The magnitude is very small, indicating that for SP500 firms investors are very sophisticated. The behavioral story accounts for small weight.

[Place Table VIII about here]

Market-to-book ratio is the market valuation of assets under firms' control. High values are often interpreted as firms being overvalued. A natural implication is that investors of firms with high market-to-book ratios are more likely to be naive and overreact to new disclosure. To examine this implication, I conduct a subsample analysis by splitting samples on firms' market-to-book ratios by fiscal year. The results summarized in Table VIII are consistent with the previous guessing. For high market-to-book ratio firms, naive investors overestimate filing period buy-and-hold excess return by 5.59 basis points, slightly higher than 5.46 for middle group and -3.21 for low group. Realized excess returns are pushed up by 0.059 basis points for high group, which is also higher than 0.058 for middle group and -0.034 for low group. As we can see, naive investors are not always overestimating. They tend to underestimate undervalued stocks.

I also conduct the subsample analysis based on size. Surprisingly, overreaction is the most sailent in medium-size firms. Large firms are the least salient. This makes sense because large firms have less information asymmetry and their disclosures are more accurate.

III. Conclusion

This study establishes a framework to quantify investors' overreaction to textual information. I solve and estimate a static model of trading with naive and sophisticated investors. Using a variant of LDA, I decompose managerial tones into topic levels. I find that overreaction to managerial tones is insignificant, pushing up the filing period excess returns by 0.03 basis points.

The results are supportive of "sophisticated" markets. Value-related information in managerial tones is learned and incorporated into prices and noises are not considered. I limit the sample to

SP500 firms which are very large. Their investors are mostly sophisticated. But the methods can be extended easily to the entire sample of public firms.

REFERENCES

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

Brown, Gregory W, Oleg Gredil, and Preetesh Kantak, 2019, Finding fortune: How do institutional investors pick asset managers?, *Available at SSRN 2797874* .

Buehlmaier, Matthias MM, and Toni M Whited, 2018, Are financial constraints priced? evidence from textual analysis, *The Review of Financial Studies* 31, 2693–2728.

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu, 2020, The structure of economic news, Technical report, National Bureau of Economic Research.

Cao, Sean, Wei Jiang, Baozhong Yang, and Alan L Zhang, 2020, How to talk when a machine is listening: Corporate disclosure in the age of ai, Technical report, National Bureau of Economic Research.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *The Journal of Finance* 75, 1371–1415.

Fedyk, Anastassia, and James Hodson, 2019, Trading on talent: Human capital and firm performance, *Available at SSRN 3017559* .

Filatov, Dmitrii, 2020, Central banks communication and the state of the economy, *Available at SSRN 3519846* .

Garcia, Diego, 2013, Sentiment during recessions, *The Journal of Finance* 68, 1267–1300.

Garcia, Diego, Xiaowen Hu, and Maximilian Rohrer, 2020, The colour of finance words, *Available at SSRN 3630898* .

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

Grundy, Bruce D, and Stefan Petry, 2020, Lda quantification of 10-k risk-factors and the information content of textual reporting, *Available at SSRN 3608594* .

Hassan, Tarek A, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, 2019, Firm-level political risk: Measurement and effects, *The Quarterly Journal of Economics* 134, 2135–2202.

Hassan, Tarek Alexander, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, 2020a, Firm-level exposure to epidemic diseases: Covid-19, sars, and h1n1, Technical report, National Bureau of Economic Research.

Hassan, Tarek Alexander, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, 2020b, The global impact of brexit uncertainty, Technical report, National Bureau of Economic Research.

Huang, Allen H, Reuven Lehavy, Amy Y Zang, and Rong Zheng, 2018, Analyst information discovery and interpretation roles: A topic modeling approach, *Management Science* 64, 2833–2855.

Jiang, Fuwei, Joshua Lee, Xiumin Martin, and Guofu Zhou, 2019, Manager sentiment and stock returns, *Journal of Financial Economics* 132, 126–149.

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu, 2019, Predicting returns with text data, Technical report, National Bureau of Economic Research.

Li, Qing, Hongyu Shan, Yuehua Tang, and Vincent Yao, 2020, Corporate climate risk: Measurements and responses, *Available at SSRN 3508497* .

Lin, Chenghua, and Yulan He, 2009, Joint sentiment/topic model for sentiment analysis, in *Proceedings of the 18th ACM conference on Information and knowledge management*, 375–384.

Lin, Chenghua, Yulan He, Richard Everson, and Stefan Ruger, 2011, Weakly supervised joint sentiment-topic detection from text, *IEEE Transactions on Knowledge and Data engineering* 24, 1134–1145.

Liu, Yukun, Jinfei Sheng, and Wanyi Wang, 2020, Do cryptocurrencies have fundamental values?, *Available at SSRN 3577208* .

Lopez-Lira, Alejandro, 2019, Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns, *Available at SSRN 3313663* .

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

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

Moghaddam, Samaneh, and Martin Ester, 2012, On the design of lda models for aspect-based opinion mining, in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 803–812.

Newman, David, Padhraic Smyth, Max Welling, and Arthur U Asuncion, 2008, Distributed inference for latent dirichlet allocation, in *Advances in neural information processing systems*, 1081–1088.

Shleifer, Andrei, and Robert W Vishny, 1997, The limits of arbitrage, *The Journal of finance* 52, 35–55.

Steyvers, Mark, and Tom Griffiths, 2007, Probabilistic topic models, *Handbook of latent semantic analysis* 427, 424–440.

Tetlock, Paul C, Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *The Journal of Finance* 63, 1437–1467.

Yang, Qinjuan, Yanghui Rao, Haoran Xie, Jiahai Wang, Fu Lee Wang, Wai Hong Chan, and E Cambria Cambria, 2019, Segment-level joint topic-sentiment model for online review analysis, *IEEE Intelligent Systems* 34, 43–50.

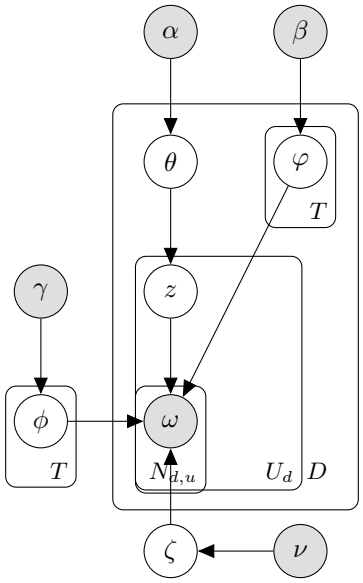
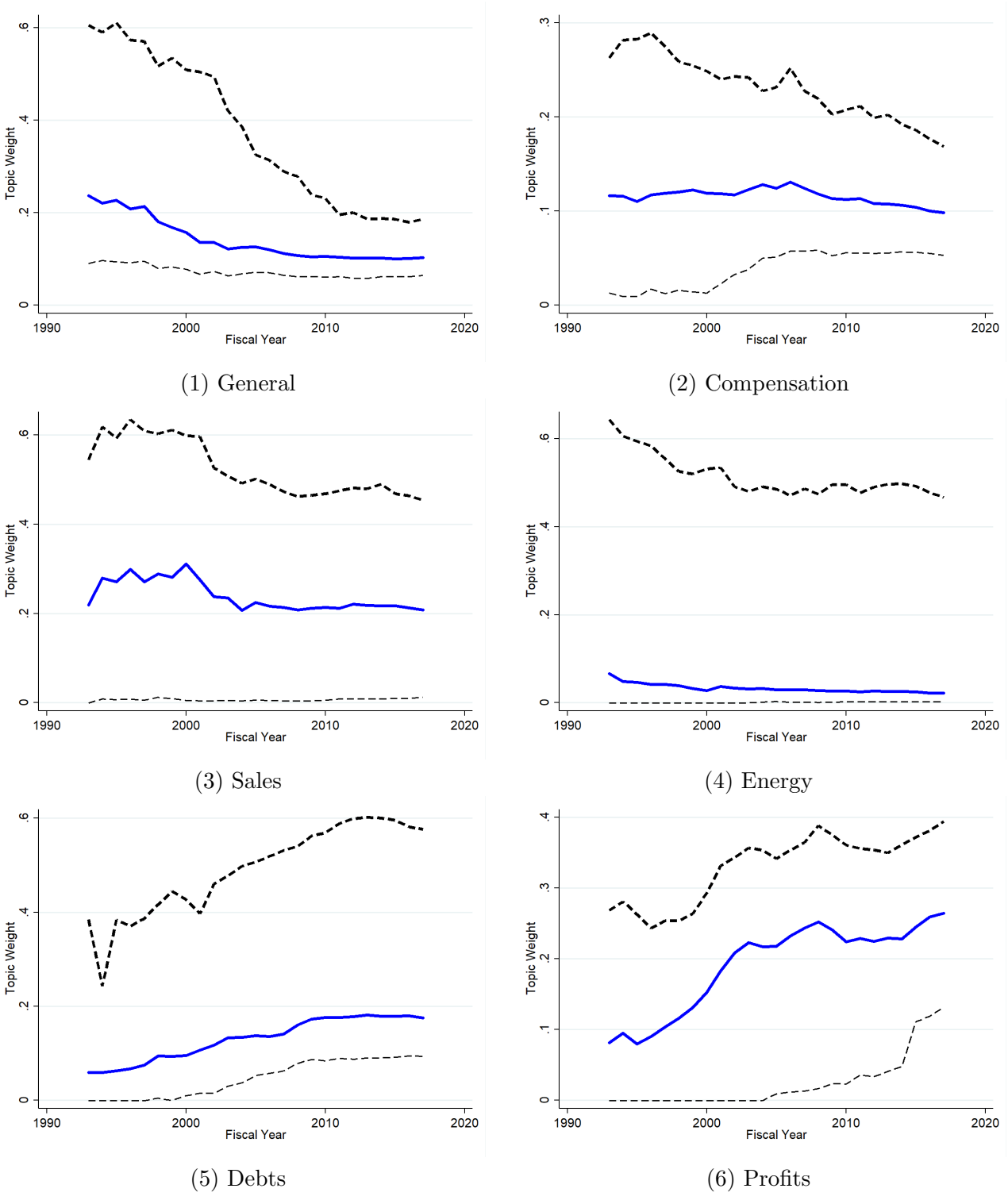


Figure 1. Plate Notation

This figure shows the bayesian network using plate notation. Shaded circles are parameters or observable variables. Unshaded circles are latent variables. Letters in triangles stand for sample sizes. Arrows stand for conditional independence. z is the topic of each sentence and ω is the token. ϕ is the distribution of each topic over neutral words and φ is the distribution over sentiment words. ζ is the probability of sentiment words. θ is the distribution over topics.



(1) General

(2) Compensation

(3) Sales

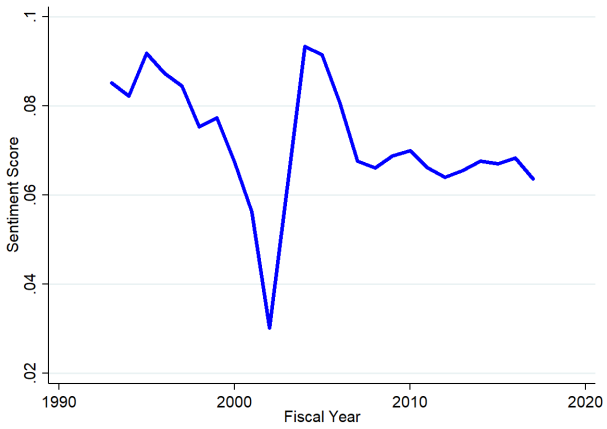
(4) Energy

(5) Debts

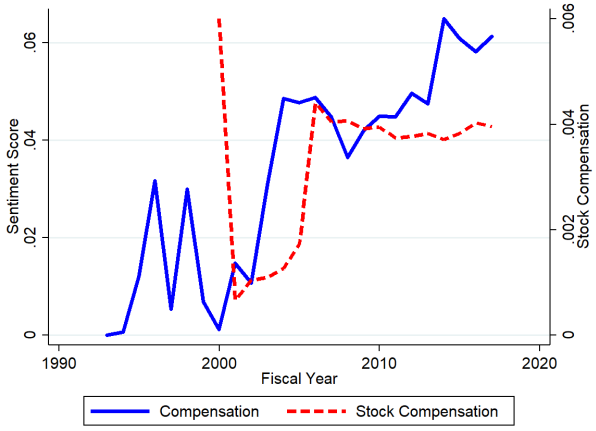
(6) Profits

Figure 2. Dynamics of Topic Weights

This figure shows the trends of topic weights. The solid lines are median across firms while the dashed lines are 5 percentile and 95 percentile.

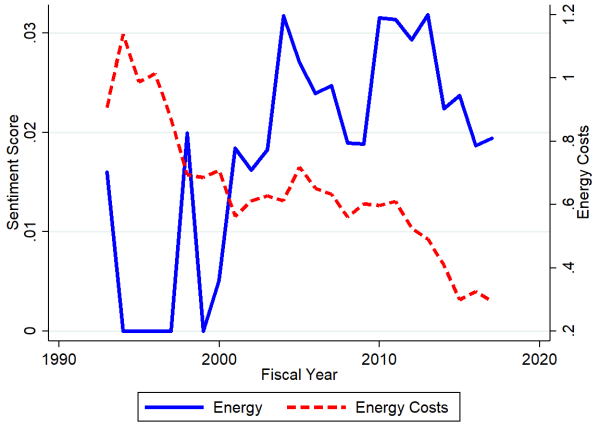


(1) General

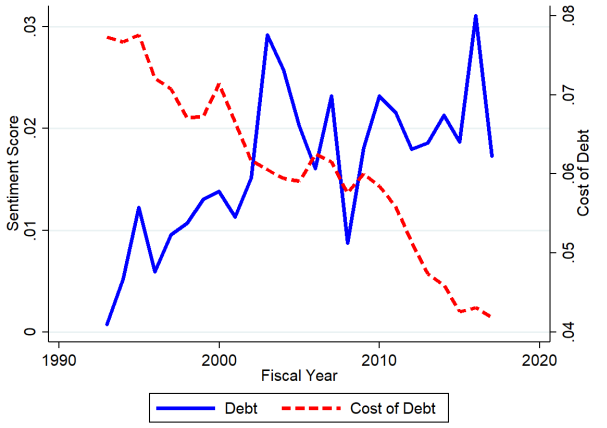


(2) Compensation

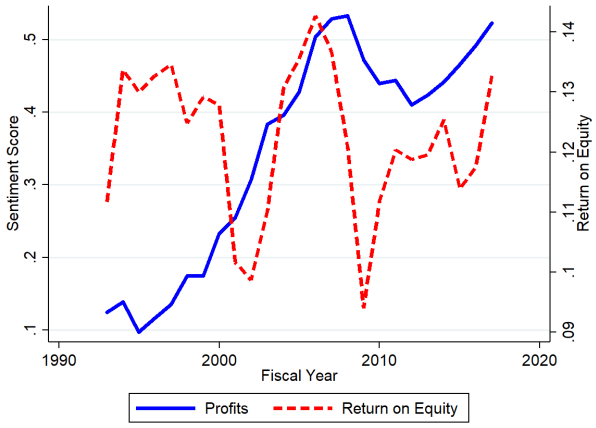
(3) Sales



(4) Energy



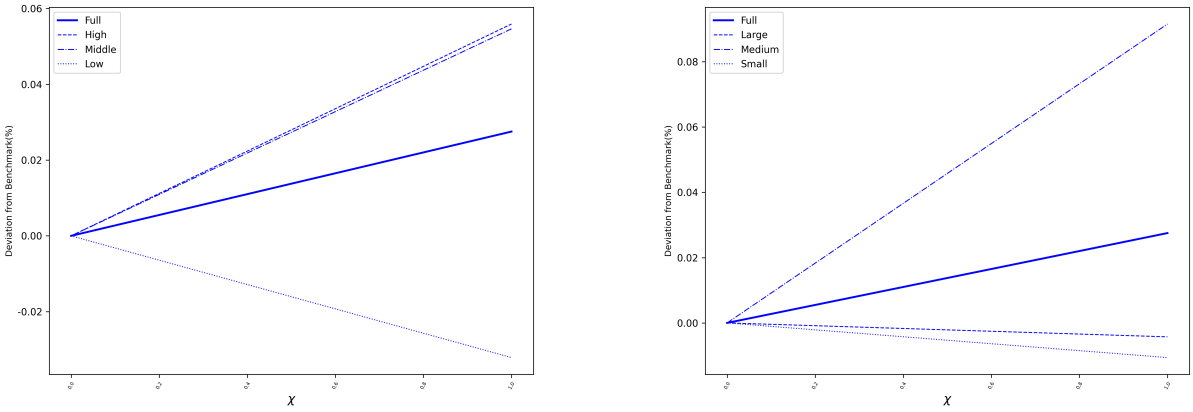
(5) Debts



(6) Profits

Figure 3. Dynamics of Topic Sentiments

This figure shows the median of topic sentiments and fundamental measures. The solid lines are sentiment scores and the dashed lines are fundamental measures. Defintion of fundamental measures are in Appendix D.



(1) Market Valuation (2) Size

Figure 4. Counterfactual Experiment

This figure shows the counterfactual experiment on different market naivity for the entire sample and different subsamples. Deviation from benchmark is defined as $E[r_t - r_{t,\chi=0}]$.

Table I 10-K Sample Creation

This table reports reports the impact of various data filters on initial 10-K sample size.

Source/Filter	Sample Size	Observations Removed
Loughran & McDonald Stage-one parse 10-K/10-K405 1993-2018 complete sample	181,420	
Include only first filing in a given year	179,614	1,806
At least 180 days between a given firm's 10-K filings	179,213	401
CRSP & COMPUSTAT match	88,768	90,445
Price on filing date minus one $\geq$ \$3	79,557	9,211
At least 60 days of returns and volume in year prior to and following file date	77,185	2,372
Book-to-market COMPUSTAT data available and book value > 0	68,807	8,378
Number of words in 10-K $\geq 2,000$	68,752	55
Remove Headers and Exhibits	63,115	5,637
Keep only firms in SP500	12,028	51,087

Table II Data and Sample

This table reports the summary statistics and correlation matrix for variables used. Document statistics are based on documents with headers and exhibits removed. Definitions of control variables are in Appendix D. Topic weights are estimates of θ. Topic sentiments are defined as $\eta^T * (\hat{\varphi}_{d,t} - I)$.

Panel A: Summary Statistics

	Mean	Std.Dev.	Min	P25	Median	P75	Max	Obs
Document Stats								
# words	37185.900	23659.870	1744.000	20017.000	35043.500	49429.000	162931.000	12022
# unqiue words	2572.348	778.433	399.000	2090.000	2668.000	3118.000	4578.000	12022
% Positive	0.718	0.199	0.103	0.586	0.707	0.846	1.435	12022
% Negative	1.576	0.566	0.214	1.214	1.557	1.907	3.550	12022
Control Variables								
Size	8.942	1.554	3.711	7.915	8.848	9.967	13.762	12022
Market-to-book ratio	2.059	1.615	0.734	1.150	1.561	2.332	34.280	11987
Leverage	0.178	0.147	0.000	0.060	0.142	0.271	0.737	11925
Stock Compensation Expense	0.007	0.010	-0.000	0.001	0.003	0.007	0.138	7809
Sales Growth	1.115	0.291	0.237	1.000	1.071	1.168	5.600	12008
Projected Energy Costs	0.917	1.205	-0.024	0.270	0.601	1.095	18.500	12022
Cost of Debt	0.070	0.089	0.006	0.044	0.059	0.074	2.760	10723
Return on Equity	0.128	0.193	-1.341	0.064	0.123	0.198	2.646	11638
Buy-and-Hold Excess Return	0.071	4.310	-28.410	-1.821	-0.027	1.859	23.734	12012
Cumulative Abnormal Return	-0.015	4.310	-29.479	-1.859	-0.077	1.778	23.008	12007
Topic Weights								
General	0.159	0.113	0.046	0.093	0.120	0.173	0.837	12022
Compensation	0.123	0.059	0.000	0.085	0.115	0.152	0.502	12022
Sales	0.232	0.167	0.000	0.073	0.228	0.350	0.735	12022
Energy	0.113	0.166	0.000	0.011	0.031	0.126	0.764	12022
debts	0.173	0.133	0.000	0.095	0.143	0.200	0.665	12022
Profits	0.198	0.108	0.000	0.138	0.211	0.277	0.477	12022
Topic Sentiment								
General	0.101	0.135	-0.117	0.036	0.070	0.120	1.631	12022
Compensation	0.040	0.183	-0.866	-0.034	0.040	0.132	1.125	12022
Sales	0.610	0.534	-0.080	0.159	0.524	0.914	3.183	12022
Energy	0.104	0.191	-0.338	-0.004	0.021	0.146	1.128	12022
debts	0.046	0.148	-0.349	-0.036	0.017	0.096	0.763	12022
Profits	0.410	0.311	-0.176	0.169	0.389	0.610	1.503	12022

Panel B: Correlation Matirx

	% Positive	% Negative	General	Compensation	Sales	Energy	debts	Profits
% Positive	1							
% Negative	0.192	1						
General	0.0768	-0.272	1					
Compensation	0.230	-0.391	0.0595	1				
Sales	0.624	-0.0816	0.0250	-0.0794	1			
Energy	-0.0665	-0.194	-0.0153	0.158	-0.383	1		
debts	0.0946	-0.0660	-0.00378	0.0929	-0.206	0.0383	1	
Profits	0.446	0.0917	-0.174	0.250	-0.122	0.0454	0.0636	1

Table III Signals Precision

This table reports signals precisions and corrected signals. Panel A reports the summary statistics of $\frac{1}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2}\frac{b_i^2}{\sigma_{\varepsilon,i}^2}$. Panel B reports correlation between $\tilde{s}_i$ and $\tilde{s}_i^*$. Topic names without asteroids refer to $\tilde{s}_i$ and those with asteroids refer to $\tilde{s}_i^*$.

Panel A: $\frac{1}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2}\frac{b_i^2}{\sigma_{\varepsilon,i}^2}$

	Mean	Std.Dev.	Min	P25	Median	P75	Max	Obs
Compensation	0.037	0.078	0.000	0.003	0.011	0.036	0.807	675
Sales	0.062	0.131	0.000	0.002	0.014	0.056	0.995	786
Energy	0.038	0.089	0.000	0.002	0.008	0.032	0.970	786
Debts	0.040	0.111	0.000	0.002	0.009	0.028	0.961	729
Profits	0.054	0.100	0.000	0.004	0.017	0.062	0.920	784

Panel B: $Corr(\tilde{s}_i, \tilde{s}_i^*)$

	Compensation	Sales	Energy	Debts	Profits
Compensation*	0.0716	0.00942	0.0163	0.0162	-0.00173
Sales*	0.0260	0.0750	-0.00111	0.0198	-0.0143
Energy*	0.0162	0.0114	0.0356	0.00655	0.00348
Debts*	-0.0146	0.00450	0.00834	0.0901	-0.00656
Profits*	0.00368	-0.00939	0.00855	-0.00941	0.0752

Table IV Main Results in Loughran and McDonald (2011)

This table reports the main results established in Loughran and McDonald (2011). Definitions of variables are in Appendix D.

	Buy-and-Hold Excess Return			CAR		
	(1)	(2)	(3)	(4)	(5)	(6)
% Negative	-0.158**	-0.153*	-0.156*	-0.114	-0.107	-0.132
	(-2.274)	(-1.739)	(-1.761)	(-1.639)	(-1.215)	(-1.487)
Size			-0.080**			-0.072**
			(-2.483)			(-2.230)
Market-to-book ratio			-0.135***			-0.208***
			(-4.410)			(-6.813)
Leverage			0.187			0.534
			(0.535)			(1.531)
Observations	12012	12012	11915	12007	12007	11910
Time FE	No	YES	YES	No	YES	YES
Fund FE	No	YES	YES	No	YES	YES

Table V Regression with Sentiment Scores

This table reports the results of regressions using raw sentiment scores. Definitions of variables are in Appendix D.

	Buy-and-Hold Excess Return			CAR		
	(1)	(2)	(3)	(4)	(5)	(6)
General	0.702**	0.633**	0.761**	0.674**	0.566*	0.694**
	(2.351)	(2.024)	(2.425)	(2.260)	(1.810)	(2.214)
Compensation	-0.159	-0.255	-0.398*	-0.151	-0.317	-0.438*
	(-0.700)	(-1.085)	(-1.663)	(-0.664)	(-1.349)	(-1.830)
Sales	-0.145*	-0.254***	-0.196**	-0.199**	-0.286***	-0.169*
	(-1.768)	(-2.651)	(-2.014)	(-2.430)	(-2.985)	(-1.741)
Energy	0.019	0.260	0.159	0.024	0.314	0.203
	(0.086)	(0.864)	(0.529)	(0.108)	(1.045)	(0.676)
debts	-0.013	0.115	0.144	-0.130	0.015	0.009
	(-0.049)	(0.361)	(0.450)	(-0.477)	(0.047)	(0.028)
Profits	0.039	0.182	0.248	0.132	0.198	0.257*
	(0.288)	(1.179)	(1.605)	(0.987)	(1.282)	(1.669)
Size			-0.106***			-0.096***
			(-3.214)			(-2.916)
Market-to-book ratio			-0.128***			-0.202***
			(-4.158)			(-6.578)
Leverage			0.096			0.459
			(0.275)			(1.315)
Observations	12012	12012	11915	12007	12007	11910
Time FE	No	YES	YES	No	YES	YES
Fund FE	No	YES	YES	No	YES	YES

Table VI OLS Regression

This table reports the results of regressions using $\tilde{s}$ & $\tilde{s}^*$. Definitions of variables are in Appendix D.

	$\tilde{s}$		$\tilde{s}^*$		$\tilde{s}$ & $\tilde{s}^*$	
	(1)	(2)	(3)	(4)	(5)	(6)
Compensation	-0.448	-0.475			-0.339	-0.403
	(-0.832)	(-0.885)			(-0.624)	(-0.743)
Sales	0.003	0.004			0.007	0.007
	(0.219)	(0.281)			(0.500)	(0.480)
Energy	0.005*	0.005*			0.005*	0.005*
	(1.766)	(1.891)			(1.736)	(1.831)
Debts	0.041	0.035			0.046	0.044
	(0.726)	(0.621)			(0.811)	(0.781)
Profits	0.005	0.004			0.001	-0.000
	(0.319)	(0.261)			(0.047)	(-0.017)
Size		-0.066*		-0.094**		-0.093**
		(-1.913)		(-2.510)		(-2.489)
Market-to-book ratio		-0.125***		-0.123***		-0.124***
		(-3.743)		(-3.360)		(-3.400)
Leverage		-0.159		0.281		0.295
		(-0.421)		(0.707)		(0.742)
General		0.971***		0.773**		0.771**
		(2.694)		(2.011)		(2.005)
Compensation*			2.121	7.986	2.487	8.642
			(0.238)	(0.832)	(0.278)	(0.898)
Sales*			-1.122**	-0.871*	-1.126**	-0.867*
			(-2.289)	(-1.714)	(-2.291)	(-1.701)
Energy*			0.112	0.085	0.109	0.081
			(1.566)	(1.170)	(1.513)	(1.120)
Debts*			-0.190	-0.253	-0.289	-0.347
			(-0.163)	(-0.214)	(-0.247)	(-0.292)
Profits*			1.088***	1.488***	1.091***	1.501***
			(2.739)	(3.566)	(2.742)	(3.590)
Observations	10354	10267	10005	9929	10005	9929
Time FE	YES	YES	YES	YES	YES	YES
Fund FE	YES	YES	YES	YES	YES	YES

Table VII Maximum Likelihood Estimation

This table reports the results of constrained maximum likelihood estimation. 90% Confidence interval is from bootstrap sampling.

Param	Estimate	90% Confidence Interval
Compensation (w_1)	-2.7989	[-17.914,12.211]
Sales (w_2)	-0.2332	[-1.099,0.737]
Energy (w_3)	0.1095	[0.017,0.217]
Debts (w_4)	-0.0191	[-1.928,2.134]
Profits (w_5)	1.0350	[0.282,2.091]
χ	0.0106	[0,0.021]
Size	-0.0629	[-0.118,-0.008]
Market-to-Book Ratio	-0.0577	[-0.169,0.046]
Leverage	0.2881	[-0.392,1.035]
General (s_i)	0.8908	[0.279,1.447]
Const	0.6748	[-0.599,1.803]
σ^2	17.2344	[16.493,18.066]

Table VIII Counterfactual Experiment

This table reports counterfactual experiments for full sample and subsamples. Return Deviation is $E[r_t - r_{t,\chi=0}]$. Valuation Deviatoin is $E\left[\sum_{i=1}^{N} w_i \frac{1}{1/\sigma_i^2+b_i^2/\sigma_{\varepsilon,i}^2} \frac{1}{\sigma_i^2}\left(\frac{s_i-a_i}{b_i}-\mu_i\right)\right]$.

Sample	Return Deviation (Basis Point)	Valuation Deviation (Basis Point)
Entire Sample	0.029	2.75
Market Valuation		
High	0.059	5.59
Middle	0.058	5.46
Low	-0.034	-3.21
Size		
Large	-0.005	-0.43
Medium	0.097	9.16
Small	-0.011	-1.06

Appendix A. Gibbs Sampling

The posterior of topic z, $P(Z|W) \approx P(W|Z)P(Z)$, where $P(W|Z)$ can be computed:

$$\begin{aligned} P(W|Z) &= \int P(W, \zeta, \varphi, \phi|Z) d\zeta d\varphi d\phi \\ &= \int P(W|Z, \zeta, \varphi, \phi) P(\zeta) P(\varphi) P(\phi) d\zeta d\varphi d\phi \end{aligned}$$

where $P(\zeta) = \prod_{d=1}^{D} P(\zeta_d)$, $P(\varphi|\beta) = \prod_{d=1}^{D} \prod_{t=1}^{T} P(\varphi_{d,t}|\beta)$, $P(\phi|\gamma) = \prod_{t=1}^{T} P(\phi_t|\gamma)$.

$$\begin{aligned} &P(W|Z, \zeta, \varphi, \phi) \\ &= \prod_{d=1}^{D} \prod_{u=1}^{U_d} \prod_{n=1}^{N_{d,u}} P(w_{d,u,n}|z_{d,u}, \zeta_d, \varphi_d, \phi_t) \\ &= \prod_{d=1}^{D} \prod_{u=1}^{U_d} \prod_{n=1}^{N_{d,u}} \prod_{t=1}^{T} \left\{ \prod_{v \in \mathcal{S}} [(\zeta_{d,t} \varphi_{d,t,v})^{s_{d,u,n,v}}]^{x_{d,u,t}} \prod_{v \in \mathcal{N}} [((1-\zeta_{d,t}) \phi_{t,v})^{s_{d,u,n,v}}]^{x_{d,u,t}} \right\} \end{aligned}$$

where $x_{d,u,t} = \mathrm{I}(z_{d,u} = t)$ and $s_{d,u,n,v} = \mathcal{I}(w_{d,u,n} = v)$. The simplifying process of the formula above is as follows:

$$\begin{aligned} &P(W|Z, \zeta, \varphi, \phi) \\ &= \prod_{d=1}^{D} \prod_{u=1}^{U_d} \prod_{n=1}^{N_{d,u}} \prod_{t=1}^{T} \zeta_{d,t}^{\mathrm{I}(w_{d,u,n} \in \mathcal{S}) x_{d,u,t}} \prod_{d=1}^{D} \prod_{u=1}^{U_d} \prod_{n=1}^{N_{d,u}} \prod_{t=1}^{T} (1-\zeta_{d,t})^{\mathrm{I}(w_{d,u,n} \in \mathcal{N}) x_{d,u,t}} \\ &* \prod_{d=1}^{D} \prod_{t=1}^{T} \prod_{v \in \mathcal{S}} (\varphi_{d,t,v})^{\sum_{u=1}^{U_d} \sum_{n=1}^{N_d} s_{d,u,n,v} x_{d,u,t}} * \prod_{t=1}^{T} \prod_{v \in \mathcal{N}} (\phi_{t,v})^{\sum_{d=1}^{D} \sum_{u=1}^{U_d} \sum_{n=1}^{N_d} s_{d,u,n,v} x_{d,u,t}} \\ &= \left(\prod_{d=1}^{D} \prod_{t=1}^{T} \zeta_{d,t}^{N_{d,t}^{\mathcal{S}}} (1-\zeta_{d,t})^{N_{d,t}^{\mathcal{N}}} \right) * \prod_{d=1}^{D} \prod_{t=1}^{T} \prod_{v \in \mathcal{S}} (\varphi_{d,t,v})^{k_s(d,t,v)} * \prod_{t=1}^{T} \prod_{v \in \mathcal{N}} (\phi_{t,v})^{k_n(t,v)} \end{aligned}$$

$N_{d,t}^{\mathcal{S}}$ is the number of sentiment words in document d assigned topic t. $N_{d,t}^{\mathcal{N}}$ is the number of neutral words in document d assigned topic t. To get the expression of $P(W|Z)$, we need to use the conjugate priors to simplify calculations. This gives us: $P(W|Z) = \left(\prod_{d=1}^{D} \prod_{t=1}^{T} \frac{B(N_{d,t}^{\mathcal{S}} + \nu, N_{d,t}^{\mathcal{N}} + \nu)}{B(\nu)} \right) * \prod_{d=1}^{D} \prod_{t=1}^{T} \left(\frac{B(K_s(d,t))}{B(\beta)} \right) * \prod_{t=1}^{T} \left(\frac{B(K_n(t))}{B(\gamma)} \right)$ where $B(.)$ is the beta function. $K_s(d,t) = \{k_s(d,t,v) : v \in \mathcal{S}\}$

and $K_n(t) = \{k_n(t,v) : v \in \mathcal{N}\}$ and the denominators $B(\beta)$ and $B(\gamma)$ have the same number of parameters as the corresponding numerators.

$$
\begin{aligned}
P(W|Z) = & \left(\prod_{d=1}^{D} \prod_{t=1}^{T} \frac{\Gamma(N_{d,t}^{\mathcal{S}} + \nu)\Gamma(N_{d,t}^{\mathcal{N}} + \nu)}{\Gamma(N_{d,t} + \nu)} \right) * \left[\frac{\Gamma(2\nu)}{\Gamma(\nu)^2} \right]^{D*T} \\
& * \left(\prod_{d=1}^{D} \prod_{t=1}^{T} \frac{\prod_{v \in \mathcal{S}} \Gamma(k_s(d,t,v) + \beta)}{\Gamma(N_{d,t}^{\mathcal{S}} + \beta * V_s)} \right) * \left[\frac{\Gamma(\beta * V_s)}{\Gamma(\beta)^{V_s}} \right]^{D*T} \\
& * \left(\prod_{t=1}^{T} \frac{\prod_{v \in \mathcal{N}} \Gamma(k_n(t,v) + \gamma)}{\Gamma(N_t^{\mathcal{S}} + \gamma * V_n)} \right) * \left[\frac{\Gamma(\gamma * V_n)}{\Gamma(\gamma)^{V_n}} \right]^{T}
\end{aligned} \tag{A1}
$$

Then we go to $P(Z) = \int P(Z|\theta)P(\theta)d\theta = \prod_{d=1}^{D} \left(\int P(z_d|\theta_d)P(\theta_d)d\theta_d \right)$

$$
P(z_d|\theta_d) = \prod_{n=1}^{U_d} P(z_{n,u}|\theta_d) = \prod_{n=1}^{U_d} \prod_{t=1}^{T} (\theta_{d,t})^{x_{d,u,t}} = \prod_{t=1}^{T} (\theta_{d,t})^{\sum_{u=1}^{U_d} x_{d,n,t}}
$$

where $x_{d,n,t} = \mathrm{I}(z_{d,u} = t)$ is an indicator variable which equals to one if the topic of uth sentence under document d is t and zero otherwise. Let's define $c(d,t) = \sum_{n=1}^{U_d} x_{d,u,t}$, i.e., the number of sentences assigned to topic t under document d. Then if $Z =< z_d | d \in \{1, \ldots, D\} >$, the likelihood is $P(Z|\theta) = \prod_{d=1}^{D} \prod_{t=1}^{T} (\theta_{d,t})^{c(d,t)}$ and the marginal distribution of Z is:

$$
\begin{aligned}
P(Z) &= \int P(Z|\theta) \prod_{d=1}^{D} \frac{1}{B(\alpha)} \prod_{t=1}^{T} (\theta_{d,t})^{\alpha - 1} d\theta \\
&= \frac{1}{(B(\alpha))^D} \int \prod_{d=1}^{D} \prod_{t=1}^{T} (\theta_{d,t})^{c(d,t) + \alpha - 1} d\theta \\
&= \frac{1}{(B(\alpha))^D} \prod_{d=1}^{D} \int \prod_{t=1}^{T} (\theta_{d,t})^{c(d,t) + \alpha - 1} d\theta_d \\
&= \left(\frac{\Gamma(T * \alpha)}{\Gamma(\alpha)^T} \right)^D \prod_{d=1}^{D} \frac{\prod_{t=1}^{T} \Gamma(c(d,t) + \alpha)}{\Gamma(U_d + \alpha * T)}
\end{aligned} \tag{A2}
$$

Following Steyvers and Griffiths (2007), Gibbs sampling approximates the joint posterior distribution $P(Z|W)$ by constructing a Markov chain. In this model, this Markov chain is constructed by

sampling the latent variable $(z_{d,u})$ given $z_{-d,-u}$, the values in other positions. For the convenience of notation, for the position $i = (d, u)$, we add a prime and use $i = (d', u')$ to distinguish from other position. First, we need to compute:

$$\frac{P(Z)}{P(Z_{-i})} = \prod_{d=1}^{D} \frac{B(C_d + \alpha)}{B(C_{d,-i} + \alpha)} = \frac{c(d', t)_{-i} + \alpha}{\sum_{t=1}^{T} c(d', t)_{-i} + T\alpha} \tag{A3}$$

$$\begin{aligned} \frac{P(W|Z)}{P(W_{-i}|Z_{-i})} &= \frac{\Gamma(N^{\mathcal{S}}_{d',t'} + \nu)}{\Gamma(N^{\mathcal{S}}_{d',t',(-i)} + \nu)} * \frac{\Gamma(N^{\mathcal{N}}_{d',t'} + \nu)}{\Gamma(N^{\mathcal{N}}_{d',t',(-i)} + \nu)} * \frac{\Gamma(N_{d',t',(-i)} + \nu)}{\Gamma(N_{d',t'} + \nu)} \\ &* \frac{\prod_{v \in \mathcal{S}_i} \Gamma(k_s(d', t, v) + \beta)}{\prod_{v \in \mathcal{S}_i} \Gamma(k_{s(-i)}(d', t, v) + \beta)} * \frac{\Gamma(N^{\mathcal{S}}_{d',t,(-i)} + \beta V_s)}{\Gamma(N^{\mathcal{S}}_{d',t} + \beta V_s)} * \frac{\prod_{v \in \mathcal{N}_i} \Gamma(k_n(t, v) + \gamma)}{\prod_{v \in \mathcal{N}_i} \Gamma(k_{n(-i)}(t, v) + \gamma)} \frac{\Gamma(N^{\mathcal{N}}_{t,(-i)} + \gamma V_n)}{\Gamma(N^{\mathcal{N}}_{t} + \gamma V_n)} \end{aligned} \tag{A4}$$

Here are some notation clarifications :$N^{\mathcal{S}}_{d',t'}$ $(N^{\mathcal{N}}_{d',t'})$ is the total number of sentiment (neutral) words in document d' assigned to topic t', $N^{\mathcal{S}}_{d',t',(-i)}$ $(N^{\mathcal{N}}_{d',t',(-i)})$ is the total number of sentiment (neutral) words in document d' if we delete the sentence i. $N_{d',t'} = N^{\mathcal{S}}_{d',t'} + N^{\mathcal{N}}_{d',t'}$. $N^{\mathcal{N}}_{t}$ is the number of neutral words in sentences whose topics are t. We will compute for $t = 1, \cdots, T$ and get $P(z_i = t|W, Z_{-i})$ with normalization.

Then we sample the topic value $z_{d,u}$ by

$$
\begin{aligned}
P(z_i = t|W, Z_{-i}) &\propto \frac{P(W|Z)}{P(W_{-i}|Z_{-i})}\frac{P(Z)}{P(Z_{-i})} \\
&\propto \frac{c(d',t)_{-i}+\alpha}{\sum_{t=1}^{T} c(d',t)_{-i}+T\alpha} \\
&* \frac{\prod_{v\in\mathcal{S}_i}\Gamma(k_s(d',t,v)+\beta)}{\prod_{v\in\mathcal{S}_i}\Gamma(k_{s(-i)}(d',t,v)+\beta)} * \frac{\Gamma(N^s_{d',t,(-i)}+\beta V_s)}{\Gamma(N^s_{d',t}+\beta V_s)} \\
&* \frac{\prod_{v\in\mathcal{N}_i}\Gamma(k_n(t,v)+\gamma)}{\prod_{v\in\mathcal{N}_i}\Gamma(k_{n(-i)}(t,v)+\gamma)}\frac{\Gamma(N^n_{t,(-i)}+\gamma V_n)}{\Gamma(N^n_t+\gamma V_n)} \\
&* \frac{\Gamma(N^{\mathcal{S}}_{d',t'}+\nu)}{\Gamma(N^{\mathcal{S}}_{d',t',(-i)}+\nu)} * \frac{\Gamma(N^{\mathcal{N}}_{d',t'}+\nu)}{\Gamma(N^{\mathcal{N}}_{d',t',(-i)}+\nu)} * \frac{\Gamma(N_{d',t',(-i)}+\nu)}{\Gamma(N_{d',t'}+\nu)}
\end{aligned} \tag{A5}
$$

The posterior distribution of parameters are the ultimate goal. Using $\zeta_{d,t}$ as an example, $P(\zeta|W) = \int P(\zeta|Z,W)P(Z|W)dZ$.

$$
\begin{aligned}
P(\zeta|Z,W) &= \int P(W|Z,\zeta,\varphi,\phi)P(\zeta) * P(\varphi)P(\phi)d\varphi d\phi \\
&\propto \prod_{d=1}^{D}\prod_{t=1}^{T}\zeta_{d,t}^{N^{\mathcal{S}}_{d,t}+\nu-1}(1-\zeta_{d,t})^{N^{\mathcal{N}}_{d,t}+\nu-1}
\end{aligned}
$$

Then we know that $\zeta_{d,t} \sim Beta(N^{\mathcal{S}}_{d,t}+\nu, N^{\mathcal{N}}_{d,t}+\nu)$. Similarly, we compute the posterior distribution for $\theta_d|(W,Z)$, $\phi_t|(W,Z)$ and $\varphi_{d,t}|(W,Z)$:

$$
\begin{aligned}
P(\theta_d|z_d) \propto P(z_d|\theta_d)P(\theta_d) &= \prod_{t=1}^{T}(\theta_{d,t})^{c(d,t)+\alpha-1} \\
P(\varphi|W,Z) &\propto \prod_{d=1}^{D}\prod_{t=1}^{T}\prod_{v\in\mathcal{S}}\varphi_{d,t,v}^{k_s(d,t,v)+\beta-1} \\
P(\phi|W,Z) &\propto \prod_{t=1}^{T}\prod_{v\in\mathcal{N}}\phi_{t,v}^{k_n(t,v)+\gamma-1}
\end{aligned}
$$

Therefore, $\theta_d|(W,Z) \sim Dir(C(d)+\alpha)$, $\varphi_{d,t} \sim Dir(K_s(d,t)+\beta)$, $\phi_t \sim Dir(K_N(t)+\gamma)$

ALGORITHM 1 (Gibbs Sampling): *Sampling process follows:*

1. *Initialisation*

   (a) *Create zero count matrix:* $D \times T \times V_S$ *matrix* Ks, $T \times V_N$ *matrix* Kn, *and* $D \times T$ *matrix* C.

   (b) *Randomly assign the value of* Z *and update count matrix*

2. *Burn-in period*

   (a) *Read one sentence* u *from document* d

   (b) *Sample a topic and a sentiment based on the probability* $P(z_i|W, Z_{-i})$ *following (A5)*

   (c) *Update the matrix* Ks, Ks, *and* C *with new sampling results.*

   (d) *Go through all sentences in all documents*

   (e) *Repeat until the iterations in burn-in period end.*

3. *Sampling period*

   (a) *Continue the sampling as in Burn-in period*

   (b) *Calculate the estimates*

   (c) *Store the estimates for each iteration*

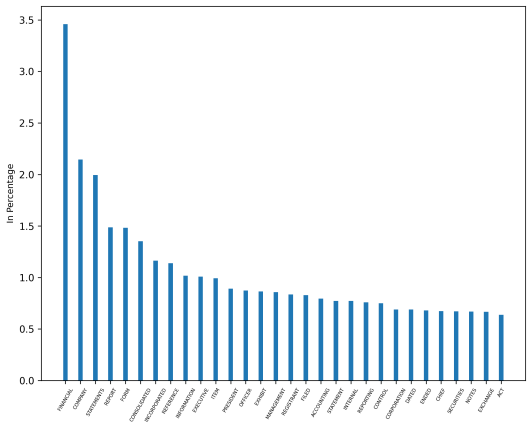
We calculate log-likelihood to determine the length of burn-in periods. We also pick up estimates every a few iterations to avoid step correlation. The log-likelihood is calculated as what follows:

Appendix B. LDA Output

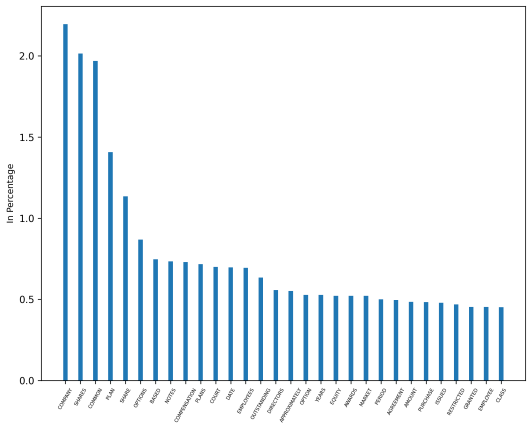
Topic Top Words:

- Topic 1 (General): FINANCIAL COMPANY STATEMENTS REPORT FORM CONSOLIDATED INCORPORATED REFERENCE INFORMATION EXECUTIVE ITEM PRESIDENT OFFICER EXHIBIT MANAGEMENT REGISTRANT FILED ACCOUNTING STATEMENT INTERNAL REPORTING CONTROL CORPORATION DATED ENDED CHIEF SECURITIES NOTES EXCHANGE ACT
- Topic 2 (Compensation): COMPANY **SHARES COMMON** PLAN **SHARE OPTIONS** BASED NOTES **COMPENSATION PLANS** COURT DATE **EMPLOYEES** OUTSTANDING **DIRECTORS** APPROXIMATELY **OPTION** YEARS **EQUITY AWARDS** MARKET PERIOD AGREEMENT AMOUNT PURCHASE ISSUED **RESTRICTED GRANTED EMPLOYEE** CLASS
- Topic 3 (Sales): **PRODUCTS** COMPANY **SERVICES BUSINESS PRODUCT CUSTOMERS** INCLUDING OPERATIONS **MARKET** SYSTEMS **SERVICE** COSTS DEVELOPMENT **REVENUE CUSTOMER MARKETS** PRIMARILY SEGMENT TECHNOLOGY MANAGEMENT INCREASED REVENUES STATES INCREASE UNITED GROWTH RELATED **OPERATING** INFORMATION ADDITION
- Topic 4 (Energy): **GAS** COSTS COMPANY **ENERGY NATURAL** OPERATIONS **OIL** APPROXIMATELY **PRODUCTION COST** OPERATING RELATED **ELECTRIC FACILITIES ENVIRONMENTAL** CUSTOMERS INCLUDING **FUEL** DUE PRIMARILY PRICES INCREASE INCREASED GENERATION RATE PROPERTIES MARKET FUTURE CAPITAL SERVICE
- Topic 5 (Debts): COMPANY **CREDIT** FINANCIAL **INTEREST RATE RISK** MARKET **DEBT TERM SECURITIES** BUSINESS **CAPITAL** BASED OPERATIONS RESULTS FUTURE **RATES** INSURANCE INVESTMENT ASSETS MANAGEMENT INCLUDING **LOANS FACILITY** INVESTMENTS RELATED CONTRACTS **FOREIGN NOTES DUE**
- Topic 6 (Profits): **INCOME TAX NET ASSETS** COMPANY FINANCIAL RELATED CONSOLIDATED STATEMENTS **EXPENSE OPERATING OPERATIONS** ACCOUNT-

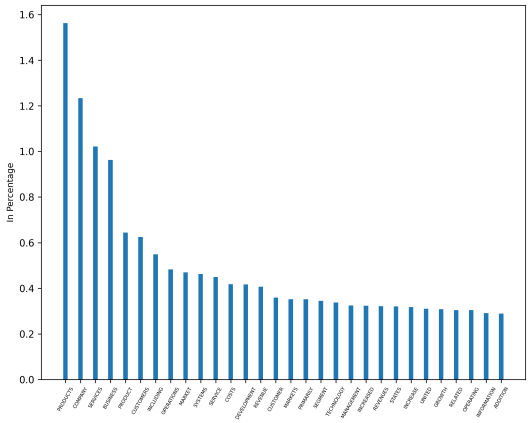
ING **COSTS** RECORDED FISCAL **INTEREST** ENDED **LIABILITIES** YEARS TABLE INCLUDED **EARNINGS DEFERRED COST TAXES** RECOGNIZED BASED **EXPENSES** AMOUNTS



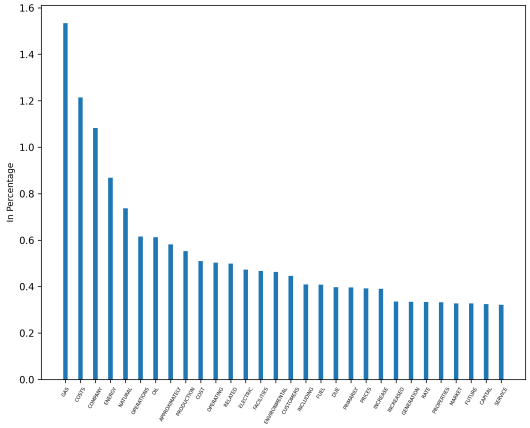
(1) General



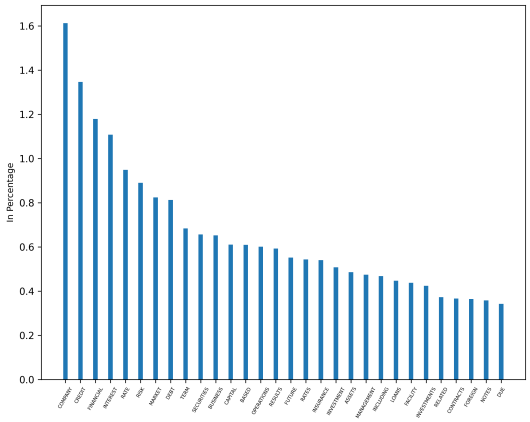
(2) Compensation



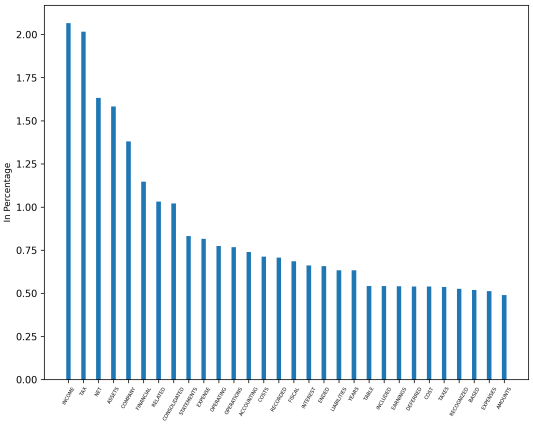
(3) Sales



(4) Energy



(5) Debts



(6) Profits

Figure 5. Topic Top Words

Appendix C. EDGAR Files

Our document files are from Loughran and McDonald Stage-One 10-X Parse files updated to 2018. I use their parsed files instead of directly downloading from EDGAR because there are some bugs. As mentioned in their paper, the EDGAR site's files are generally reliable but some files was corrupted. They validate the downloads using past versions of the data whenever possible to avoid some of these errors. Overall, the "Stage One Parse" cleans each filing document of extraneous materials. A substantial portion of an EDGAR text filing's content consists of HTML code, embedded PDF's, jpg's and other artifacts not typically of interest. The Stage One Parse excludes markup tags, ASCII-encoded graphics, and tables. They insert their own markup tags within a header at the beginning of the compressed document and tags to delineate all exhibits in the document.

I restrict our sample to documents filed from 2011 to 2018 by SP500 constituents. Large firms have big reputation's costs and their documents will disclose more information. Informative documents avoids noise and improves results. SP500 constituents are from CRSP Monthly SP500 Index database. "Stage One Parse" uses CIK as the firm identifier. CRSP monthly SP500 Index database uses PERMNO as the identifier. Stage One Parse produces 266,174 files in the sample period while Loughran and McDonald 10X File Summaries file provided here has 260,425 records in this period.

I perform a set of standard cleaning procedures following Garcia et al. (2020). I first eliminate all number characters, punctuation, and anything that are not alphanumeric characters. Then I remove single character words and those in all stopword lists provided by Loughran and McDonald. I include a handful of terms into this stopword list. They are

- 'january', 'february', 'march', 'april', 'may', 'june', 'july', 'august', 'september', 'october', 'november', 'december', 'month', 'year'
- 'money', 'million', 'millions', 'thousand', 'thousands', 'hundred', 'hundreds', 'ten', 'billion', 'billions', 'tillion','trillions'
- 'per', 'non', 'pre', 'may', 'can', 'might', 'could', 'will', 'would', 'also', 'see','saw','percent'

Appendix D. Variable Definitions

Size is the logarithm of asset book value (Compustat item AT), deflated by PPI.

Market-to-book ratio is defined as (AT + PRCC_F*CSHO - SEQ - TEDB)/AT.

Leverage is defined as (DLTT + DLC)/(AT + PRCC_F*CSHO - SEQ - TXDB) .

Stock Compensation Expense is defined as STKCO/AT.

Sales Growth is defined as SALE divided by lagged SALE.

Projected Energy Costs is calculated in the following way. First, project the total operating cost, Compustat Item XOPR, on the global price of energy index using linear regressions and then scale it with book value of assets, COMPUSTAT item AT. The time series of global price of energy index is available in Federal Reserve's database [5].

Cost of Debt is defined as Item XINT/(Item DLTT + Item DLC).

Return on Equity is defined as IB/((BE+LAG(BE)/2)).

Buy-and-Hold Excess Return is defined as the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return from days 0 through 3.

Cumulative Abnormal Return is defined as the sum of abnormal returns from days 0 through 3 using CAPM to estimate beta. Estimation window is 110 days with a gap of 50 days between the end of estimation window and event dates.

Appendix E. Gibbs Sampling

I start with getting the posterior distribution, $P(W, L, Z|\zeta)$. The posterior distribution of θ_d conditional on latent topics z_d is:

$$P(\theta_d|z_d) \propto P(z_d|\theta_d)P(\theta_d|\alpha) \propto \prod_{t=1}^{T}(\theta_{d,t})^{\alpha-1}\prod_{t=1}^{T}(\theta_{d,t})^{c(d,t)}$$
$$\propto \prod_{t=1}^{T}(\theta_{d,t})^{c(d,t)+\alpha=1} \tag{E1}$$

[5] https://fred.stlouisfed.org/series/PNRGINDEXM

Therefore, $\theta_d|z_d \sim Dir(C_d + \alpha)$ where $C_d =< c(d,t)|t \in \{1,\ldots,T\} >$. Given the topic mixture θ_d in document d, the likelihood of latent topics in document d can be calculated:

$$P(z_d|\theta_d) = \prod_{n=1}^{U_d} P(z_{n,u}|\theta_d) = \prod_{n=1}^{U_d}\prod_{t=1}^{T}(\theta_{d,t})^{x_{d,u,t}} = \prod_{t=1}^{T}(\theta_{d,t})^{\sum_{u=1}^{U_d} x_{d,n,t}} \tag{E2}$$

where $x_{d,n,t} = \mathrm{I}(z_{d,u} = t)$ is an indicator variable which equals to one if the topic of uth sentence under document d is t and zero otherwise. Let's define $c(d,t) = \sum_{n=1}^{U_d} x_{d,u,t}$, i.e., the number of sentences assigned to topic t under document d. Then if $Z =< z_d|d \in \{1,\ldots,D\} >$, the likelihood is $P(Z|\theta) = \prod_{d=1}^{D}\prod_{t=1}^{T}(\theta_{d,t})^{c(d,t)}$ and the marginal distribution of Z is:

$$\begin{aligned}
P(Z) &= \int P(Z|\theta)\prod_{d=1}^{D}\frac{1}{B(\alpha)}\prod_{t=1}^{T}(\theta_{d,t})^{\alpha-1}d\theta \\
&= \frac{1}{(B(\alpha))^D}\int\prod_{d=1}^{D}\prod_{t=1}^{T}(\theta_{d,t})^{c(d,t)+\alpha-1}d\theta \\
&= \frac{1}{(B(\alpha))^D}\prod_{d=1}^{D}\int\prod_{t=1}^{T}(\theta_{d,t})^{c(d,t)+\alpha-1}d\theta_d \\
&= \left(\frac{\Gamma(T*\alpha)}{\Gamma(\alpha)^T}\right)^D\prod_{d=1}^{D}\frac{\prod_{t=1}^{T}\Gamma(c(d,t)+\alpha)}{\Gamma(U_d+\alpha*T)}
\end{aligned} \tag{E3}$$

The sentiment mixture under document d, π_d, follows $Dir(\gamma)$ and thus $P(\pi_{d,t}|\gamma) = 1/B(\gamma) * \prod_{s=1}^{S}(\pi_{d,t,s})^{\gamma-1}$. The sentence-level sentiment depends on topic and sentiment mixture in that document. $P(l_{d,u}|z_{d,u},\pi_d) = \prod_{t=1}^{T}\left[\prod_{s=1}^{S}(\pi_{d,t,s})^{y_{d,u,s}}\right]^{x_{d,u,t}}$, where $y_{d,u,s} = \mathrm{I}(l_{d,u} = s)$ is an indicator variable which equals to one if the sentence sentiment is s and zero otherwise. The likelihood of l_d is

$$P(l_d|z_d,\pi_d) = \prod_{t=1}^{T}\prod_{s=1}^{S}(\pi_{d,t,s})^{\sum_{u=1}^{U_d} x_{d,u,t}y_{d,u,s}} \tag{E4}$$

Let's define $h(d,t,s) = \sum_{u=1}^{U_d} x_{d,u,t}y_{d,u,s}$, i.e., the number of sentence in document assigned to topic

t and sentiment s. Integration over π_d yields $P(l_d|z_d)$:

$$\begin{aligned} P(l_d|z_d) &= \int P(l_d|z_d,\pi_d)d\pi_d \\ &= \int \prod_{t=1}^{T}\prod_{s=1}^{S}(\pi_{d,t,s})^{h(d,t,s)} * \prod_{t=1}^{T} 1/B(\gamma) * \prod_{s=1}^{S}(\pi_{d,t,s})^{\gamma-1}d\pi \\ &= \frac{\prod_{t=1}^{T} B(h(d,t)+\beta)}{B(\gamma)^T} \end{aligned} \tag{E5}$$

And $P(L|Z) = \prod_{d=1}^{D} P(l_d|z_d) = \left(\frac{\Gamma(S*\gamma)}{\Gamma(\gamma)^S}\right)^{T*D} \prod_{d=1}^{D}\prod_{t=1}^{T} \frac{\prod_{s=1}^{S}\Gamma(h(d,t,s)+\beta)}{\Gamma(U_{d,t}+\beta*S)}$. where $U_{d,t}$ is the number of sentences assigned to topic t.

Finally, generative process defines the likelihood of words $\omega_{d,u,n}$. First, randomly decide whether the word belongs to $\mathcal{S}$ or $\mathcal{N}$. Second, if given the topic and sentiment of this sentence, randomly choose from either $\mathcal{S}$ or $\mathcal{N}$.

$$\begin{aligned} P(w_{d,u,n}|z_{d,u},l_{d,u},\varphi,\phi,\zeta) = &\left[\prod_{t=1}^{T}\prod_{s=1}^{S}\left[\prod_{v=1}^{V_S}(\zeta\varphi_{t,s,v})^{s_{d,u,n,v}}\right]^{y_{d,u,s}}\right]^{x_{d,u,t}} * \\ &\left[\prod_{t=1}^{T}\prod_{v=1}^{V_N}((1-\zeta)\phi_{t,v})^{s_{d,u,n,v}}\right]^{x_{d,u,t}} \end{aligned} \tag{E6}$$

where $s_{d,u,n,v} = \mathcal{I}(\omega_{d,u,n} = v)$. The likelihood function for the whole corpora is:

$$\begin{aligned} \prod_{d=1}^{D} P(w_d|z_d,l_d,\varphi,\phi,\zeta) &= \prod_{d=1}^{D}\prod_{u=1}^{U_d}\prod_{n=1}^{N_{d,u}} P(w_{d,u,n}|z_{d,u},l_{d,u},\varphi,\phi,\zeta) \\ &= \zeta^{\#sentiment}(1-\zeta)^{\#neutral}\prod_{t=1}^{T}\prod_{s=1}^{S}\prod_{v=1}^{V_S}\varphi_{t,s,v}^{k^S(t,s,v)} * \prod_{t=1}^{T}\prod_{v=V_S+1}^{V}\phi_{t,v}^{k^N(t,v)} \end{aligned} \tag{E7}$$

Still I define $k^S(t,s,v) = \sum_{d=1}^{D}\sum_{u=1}^{U_d}\sum_{n=1}^{N_{d,u}} s_{d,u,n,v}y_{d,u,s}x_{d,u,t}$, i.e., the number of times that sentiment-charged word v in the corpora that belongs to topic t and sentiment s, and $k^N(t,v) = \sum_{d=1}^{D}\sum_{u=1}^{U_d}\sum_{n=1}^{N_{d,u}} s_{d,u,n,v}x_{d,u,t}$, the number of times that the sentiment-neutral word v in the corpora that belongs to topic t. And $\#sentiment$ is the sentiment-charged word count, and $\#neutral$, which is the sentiment-neutral word count.

Given the priors of ϕ and φ, $P(\phi_t|\eta) = \frac{1}{B(\eta)}\prod_{v=1}^{V_S}(\phi_{t,v})^{\eta-1}$ and $P(\varphi_{t,s}|\beta) = \frac{1}{B(\beta)}\prod_{v=1}^{V_N}(\varphi_{t,s,v})^{\beta-1}$.

Integrating over φ and ϕ gives:

$$
\begin{aligned}
P(W|Z,L,\zeta) &= \int P(W|Z,L,\varphi,\phi)P(\varphi|\beta)P(\phi|\eta)d\varphi d\phi \\
&= \frac{\prod_{t=1}^{T}\prod_{s=1}^{S} B(K^S(t,s)+\beta)\prod_{t=1}^{T} B(K^N(t)+\eta)}{B(\beta)^{S*T}B(\eta)^T} * \zeta^{\#sentiment}(1-\zeta)^{\#neutral} \\
&= \left(\frac{\Gamma(\beta * V_S)}{\Gamma(\beta)^{V_S}}\right)^{S*T} \prod_{t=1}^{T}\prod_{s=1}^{S} \frac{\prod_{v=1}^{V_S}\Gamma(k^S(t,s,v)+\beta)}{\Gamma(N_{t,s}^S+\beta * V_S)} \\
&* \left(\frac{\Gamma(\eta * V_N)}{\Gamma(\eta)^{V_N}}\right)^{T} \prod_{t=1}^{T} \frac{\prod_{v=V_S+1}^{V}\Gamma(k^N(t,v)+\eta)}{\Gamma(N_t^N+V_N*\eta)} * \zeta^{\#sentiment}(1-\zeta)^{\#neutral}
\end{aligned} \tag{E8}
$$

where $N_{t,s}^S$ is the number of sentiment-charged words assigned to topic t and sentiment s and N_t^N is the number of sentiment-neutral words assigned to topic t.

Finally, the joint distribution of (W,Z,L) is

$$
P(W,Z,L|\zeta) = P(W|Z,L,\zeta) * P(L|Z) * P(Z) \tag{E9}
$$

For the convenience of notation, for the position $i=(d,u)$, I add a prime and use $i=(d',u')$ to distinguish from other position. And the values of associated latent variables are denoted as d

$$
P(z_i,l_i|W,Z_{-i},L_{-i}) \propto \frac{P(W|Z,L)}{P(W_{-i}|Z_{-i},L_{-i})}\frac{P(L|Z)}{P(L_{-i}|Z_{-i})}\frac{P(Z)}{P(Z_{-i})}
$$

Then let's calculate $\frac{P(W|Z,L)}{P(W_{-i}|Z_{-i},L_{-i})}$, $\frac{P(L|Z)}{P(L_{-i}|Z_{-i})}$, and $\frac{P(Z)}{P(Z_{-i})}$.

$$\begin{aligned}\frac{P(L|Z)}{P(L_{-i}|Z_{-i})} &= \frac{B(H(d',t')+\gamma)}{B(H(d',t')_{-i}+\gamma} \\ &= \frac{\prod_{s=1}^{S}\Gamma(h(d',t',s)+\gamma)}{\Gamma(\sum_{s=1}^{S}h(d',t',s)+S\gamma)}\frac{\Gamma(\sum_{s=1}^{S}h(d',t',s)+S\gamma)}{\prod_{s=1}^{S}\Gamma(h(d',t',s)+\gamma)} \\ &= \frac{h(d',t',s')_{-i}+\gamma}{\sum_{s=1}^{S}h(d',t',s)_{-i}+S\gamma}\end{aligned} \tag{E10}$$

$$\begin{aligned}\frac{P(Z)}{P(Z_{-i})} &= \prod_{d=1}^{D}\frac{B(C_d+\alpha)}{B(C_{d,-i}+\alpha)} \\ &= \frac{c(d',t')_{-i}+\alpha}{\sum_{t=1}^{T}c(d',t)_{-i}+T\alpha}\end{aligned} \tag{E11}$$

The first two terms will be pretty much the same as in LDA or JST. The calculation is just from the word level to sentence level. The third term will be more complicated since we have to compute the word-level likelihood on a sentence level and hence all the words in the sentence will be accounted for.

$$\begin{aligned}\frac{P(W|Z,L)}{P(W_{-i}|Z_{-i},L_{-i})} &= \frac{\prod_{v=1}^{V_S}[k^S(t',s',v)_{-i}+\beta+\sum_{n=1}^{N_{d,u}}\mathcal{I}(\omega_{d,u,n}=v)-1]^{\mathcal{I}(v\in Sent_{d',u'})}}{\prod_{n=1}^{N^S_{d,u}}[\sum_{v=1}^{V_S}k^S(t',s',v)_{-i}+\beta V_S+N^S_{d,u}-n]} \\ &* \frac{\prod_{v=1}^{V_N}[k^N(t',v)_{-i}+\eta+\sum_{n=1}^{N_{d,u}}\mathcal{I}(\omega_{d,u,n}=v)-1]^{\mathcal{I}(v\in Sent_{d',u'})}}{\prod_{n=1}^{N^N_{d,u}}[\sum_{v=1}^{V_N}k^N(t',v)_{-i}+\eta V_N+N^N_{d,u}-k]} \\ &* \zeta^{N^S_{d,u}}(1-\zeta)^{N^N_{d,u}}\end{aligned} \tag{E12}$$

Combine the three parts and $P(z_i,l_i|W,Z_{-i},L_{-i})$ follows multinomial distribution.

Gibbs sampling algorithm is proposed in Proposition 3. Gibbs sampling approximates the joint distribution by constructing a Markov chain. In this model, this Markov chain is constructed by

sampling the latent variable $(z_{d,n}, l_{d,n})$ given the values of (z, l) in other positions. Notice that in my model the sentiment and topic are on the sentence level and I assume that one sentence has only one topic. This means that all the words in the sentence are from the same topic and the same sentiment. This structure determines that the clustering relies on sentence-level word co-occurrence and links the sentiment to the topic by physical approximation. Yang, Rao, Xie, Wang, Wang, Chan, and Cambria (2019) propose that this structure can be on segment level, i.e, the clause of a sentence. My model can be easily extended to segment level by replacing the sentence with segment.

PROPOSITION 3 (Gibbs Sampling): *Sampling process follows:*

1. *Initialisation*
   (a) *Create zero count matrix:* $T \times S \times V_S$ *matrix* KS, $T \times V_N$ *matrix* KN, $D \times T \times S$ *matrix* H, *and* $D \times T$ *matrix* C.
   (b) *Randomly assign the value of* Z *and* L *and update count matrix*
2. *Burn-in period*
   (a) *Read one sentence* u *from document* d
   (b) *Sample a topic and a sentiment based on the probability* $P(z_i, l_i | W, Z_{-i}, L_{-i})$. *The expression is given in E*
   (c) *Update the matrix* KS, KN, H, *and* C *with new sampling results.*
   (d) *Go through all sentences in all documents have been processed.*
   (e) *Repeat until the iterations in burn-in period end.*
3. *Sampling period*
   (a) *Continue the sampling as in Burn-in period*
   (b) *Calculate the estimates based on corollary .*
   (c) *Each five iterations of going through all documents, store the estimates*
4. *Final parameter estimates are the mean of estimates stored in sampling period.*

Even though generating accurate distribution in the long run, it is hard to judge whether the sampling steps converge. We need to rely on log-likelihood and after the burn-in periods we need to pick up samples every a few steps to avoid step correlation.

Appendix F. Proof of Proposition 2

As proved in E, $p(z_d) = \int p(z_d|\theta_d)p(\theta_d)d\theta = \frac{B(C_d+\alpha)}{B(\alpha)}$ and $p(\pi_d) = \frac{1}{B(\alpha)^T}\prod_{t=1}^{T}\prod_{s=1}^{S}(\pi_{d,t,s})^{\gamma-1}$. Then The posterior distribution of π_d conditional on latent topics z_d and sentiment l_d is:

$$P(\pi_d|z_d, l_d) \propto P(l_d|z_d, \pi_d)P(z_d)P(\pi_d) \propto \prod_{t=1}^{T}\prod_{s=1}^{S}(\pi_{d,t,s})^{h(d,t,s)+\gamma-1} \tag{F1}$$

Therefore, $\pi_{d,t}|z_d, l_d \sim Dir(H(d,t)+\gamma)$ where $H(d,t) =< h(d,t,s)|s \in \{1,\ldots,S\} >$

The posterior distribution of $\varphi_{s,t}$ based on W, Z, and L:

$$\begin{aligned} P(\varphi|W,Z,L) &\propto \int P(W,Z,L,\varphi,\phi)d\phi \propto \int P(W|Z,L,\varphi,\phi)P(\varphi)P(\phi)d\phi \\ &\propto \prod_{t=1}^{T}\prod_{s=1}^{S}\prod_{v=1}^{V_S}(\varphi_{t,s,v})^{k^S(t,s,v)+\beta-1} \end{aligned} \tag{F2}$$

Therefore $\varphi_{t,s}|W,Z,L \sim Dir(K^S(t,s)+\beta)$. Similarly,

$$\begin{aligned} P(\phi|W,Z,L) &\propto \int P(W,Z,L,\varphi,\phi)d\varphi \propto \int P(W|Z,L,\varphi,\phi)P(\varphi)P(\phi)d\varphi \\ &\propto \prod_{t=1}^{T}\prod_{v=1}^{V_N}(\phi_{t,v})^{k^N(t,v)+\eta-1} \end{aligned} \tag{F3}$$

Therefore $\phi_t|W,Z,L \sim Dir(K^N(t)+\eta)$.

Appendix G. Approximate Distributed Algorithm

The algorithm is slow in execution when training sample is large. The intrinsic structure of MCMC makes it difficult to distribute on HPC. For the sake of acceleration, we need to develop a variant of this model, which can approximate the original and can be distributed on clusters. Our approximated model follows the idea of AD-LDA first presented in Newman, Smyth, Welling, and Asuncion (2008). The Gibbs sample process depends on four statistics: $h(d,t,s)$, $c(d,t)$, $k^S(t,s,v)$, and $k^N(t,v)$. $k^S(t,s,v)$ and $k^N(t,v)$ are global statistics, i.e., built on information from all documents while $h(d,t,s)$ and $c(d,t)$ are only reliant on information of document t. We distribute documents on q multiple processors, $1,\cdots,q$, execute Gibbs sampling in each processor,

and update globally. Gibbs sampling rule is summarized in $P(z_i, l_i|W, Z_{-i}, L_{-i})$, which depends on $c(d,t)$, $h(d,t,s)$, $k^N(t,v)$, and $k^s(t,s,v)$. The sequential essence of MCMC will create some problems for the calculation of $k^N(t,v)$ and $k^s(t,s,v)$ when distributed. The global updating rule is:

$$k^s(t,s,v) \leftarrow k^s(t,s,v) + \sum_q [k_q^s(t,s,v) - k^s(t,s,v)]$$

$$k^N(t,v) \leftarrow k^N(t,v) + \sum_q [k_q^N(t,v) - k^N(t,v)]$$

$$k_q^s(t,s,v) \leftarrow k^s(t,s,v); k_q^N(t,v) \leftarrow k^N(t,v)$$

In the beginning of each round of sampling, k^s and k^N are the same for each processor. One round of sampling will generate q different matrices, k_q^s and k_q^N. Before the next round of sampling on q processors, we update k_q^s and k_q^N on each processor based on the rules above and make them equal again.