Investor Inattention, Financial Narrative, and Tone-Based Heuristics

Miles Gietzmann[1,2], Francesco Grossetti[1,2], and Craig Lewis[3]

[1]*Department of Accounting, Bocconi University, Milan, Italy*
[2]*Bocconi Institute for Data Science and Analytics, Milan, Italy*
[3]*Owen Graduate School of Management, Vanderbilt University, Nashville, USA*

November 5, 2021
* * * Preliminary and Incomplete * * *
* * * Do not circulate this manuscript without authors permission * * *

Abstract

Keywords: investor inattention, analyst coverage, textual analysis, machine learning

Abstract

Typically, financial communication combines a mix of quantitative information and qualitative narrative disclosure. In such settings, prior research has shown that introducing measures of tone (sentiment) can help investors improve inferences about firm valuation. At issue is whether tonal summary measures provide sufficient information or instead are simply timely heuristics that help direct initial reactions from temporarily inattentive investors. If it is the later, this provides an institutional description consistent with recent theoretical work on pricing errors. That is, pricing errors can occur when investors are inundated with information (seemingly leading to inattention) for instance during the annual earnings season, and rather than immediately conduct detailed analyses of all (slow to process) qualitative data, quickly react to sentiment and summary analyst reports. We test hypotheses that posit sentiment measures are temporarily important but that eventually investors go beyond the tone of narrative and dig deeper into other qualitative textual features such as the use of optimism. In order to improve identification, we exploit several institutional differences in our grand sample of firms. Firms differ in whether they are followed (reported on) by analysts which allows us to test whether investors can temporarily free ride on analyst reports when available. Next, we exploit

the fact that some firms use staggered reporting in which they initially report headline numbers and then subsequently, release full (lengthy) 10K filing which include detailed narrative text in the Management Discussion and Analysis (MD&A) statement. Thirdly we consider both short and long window reactions in order to try to detect a delayed reaction to slower to process higher order qualitative dimensions of text. Rather than use a dictionary-based approach this research uses a machine learning (Naive Bayes Classifier) algorithm to identify the narrative attributes of text that uses a novel survey approach to crowd source classification.

1 Introduction

When management of listed companies add narrative to financial filings they typically take advice from investor relations professionals [Bushee and Miller, 2012]. Such advice typically aims at putting companies in the best possible light. For instance, a CEO could be advised to be very upbeat in the narrative even if quantitative financials are not particularly supportive.[1] Given that narrative content may be strategically managed, and possibly not closely related to objective facts, rational investors may de-emphasize their reliance on certain aspects of such content [Huang et al., 2014]. For example, if a firm consistently provides upbeat narrative, it may be rational to discount this characteristic when analyzing company disclosures.[2]

However, given the inherent complexity of the information contained in financial filings, investors with limited attention may adopt simple heuristics [Tversky and Kahneman, 1973] that help them process the massive amount of data they contain [Huang et al., 2014].[3] When investors face time constraints that prevent them from fully processing narrative content or they have an aversion to processing technical content, inattentive investors may act on underlying decision biases by adopting heuristic algorithms that focus on salient content [Hirshleifer and Teoh, 2003]. One advantage of heuristic algorithms is that they allow investors to make rapid assessments of complex content, even if they come at the expense of ignoring potentially valuable information. Unfortunately, investor inattention can manifest itself in the form of pricing errors. Recent theoretical work [Hendershott et al., 2021] shows that delayed reactions to available information by inattentive investors does in fact influence share price - a necessary condition for any test of the inattentive investor hypothesis.

Following the lead of Huang et al. [2014], we explore the possibility that inattentive investors initially evaluate narrative content using tone-based algorithms.[4] Although there is no direct evidence that investors actually use tone-based heuristics, given the emphasis on narrative tone in the academic literature and by institutional quantitative investors, we con-

[1] Not surprisingly, an appreciation of the strategic choice of tone is well known in other settings such as political science [Grimmer and Stewart, 2013].

[2] Alternatively, management may choose to use excessively negative tone to strategically fend off future litigation.

[3] Hirshleifer and Teoh [2003] discuss the theory and evidence as it relates to decision biases that are applicable to investor inattention. They describe a number of decision biases that are particularly relevant in the context of narrative content. These include: effort constraints [Kahneman, 1973], strongly worded content, or "salience" [Fiske and Taylor, 1991], a tendency to under-weight abstract and technical narrative [Kahneman and Tversky, 1973; Nisbett and Ross, 1980], and the so-called *availability heuristic* [Tversky and Kahneman, 1973]. The availability heuristic posits that investors assess the likelihood that narrative content contains a specific message according to their ability to retrieve confirmatory examples from memory.

[4] Huang et al. [2014] use a dictionary-based approach to measure abnormal tone in earnings press releases to examine "tone management". They find that abnormal tone (estimated as the residual from a regression model) has a positive stock return effect and a delayed negative reaction in the two quarters afterward.

sider a tone-based heuristic to be a natural choice. Investors can rapidly determine overall tone from narrative content in a way that does not require a careful reading of the document. For example, institutional investors commonly employ machine learning to infer sentiment, which is then incorporated into quantitative trading strategies. The use of heuristics would be especially appealing to investors that face time constraints or do not have the analytical skills needed to fully process financial filings.

A second factor that may also contribute to investor inattention is the opportunity to free-ride on the information produced by equity research analysts.[5] Since equity analysts are evaluated, at least in part, on the quality of the information they provide, inattentive investors are likely to view them as expert information processors. When investors use analyst research in conjunction with tone-based heuristics, they effectively delegate much of the information collection process to analysts. This can be problematic because analysts tend to rapidly update forecasts immediately following earnings announcements and there is no guarantee that any in-depth research, implicitly expected by investors, is performed [Guo et al., 2020]. When this happens, share prices would not reflect all public information because neither investors nor analysts would have fully processed it.

This gives rise to our first question. How do investors react to narrative content when firms do not have analyst coverage? In the absence of analyst coverage, investors do not have an opportunity to free-ride on the information produced by analysts. While it may be the case that investors continue to use tone-based heuristics to form partial assessments of the narrative contained in financial filings, the absence of analysts will force investors to perform independent analysis.

An example of the type of additional independent analysis that investors perform relates to narrative content. While it is clear from the empirical evidence that tone is an undeniably important aspect of a firm's narrative content [Muslu et al., 2014], it is not the only dimension one can look at. This raises another interesting question. Is an estimate of tone a sufficient metric to process narrative content? In other words, do investors consider additional linguistic attributes in addition to tone when evaluating financial narrative?

One prediction of the inattentive investor hypothesis [Hirshleifer and Teoh, 2003] we consider in this paper is that underreaction is more pronounced for firms that have analyst coverage. We predict that investors are less likely to fully evaluate linguistic attributes because they

[5] Free-riding can take one of two forms. First, coverage firms typically host earnings calls that discuss current results with analysts who are permitted to ask questions that provide additional incremental information. The information conveyed may subsume some of the information contained in 10-K (and 10-Q) reports. Second, analysts may rapidly update earnings forecasts and research reports to reflect this new information. An inattentive investor may simply wait until the analyst updates are rel;eased before adjusting their views on firm value.

are able to delegate information processing to analysts. By contrast, investors in non-coverage firms will be more likely to evaluate the complete narrative because they are responsible for performing their own due diligence. This is expected to include an evaluation of all relevant narrative attributes, including for example optimism. Consistent with this prediction, Muslu et al. [2014] show that stock returns are associated with tone and optimism. We extend Muslu et al. [2014] by framing our empirical tests in the context of the inattentive investor hypothesis. This provides additional economic insights into the manner in which investors process narrative content.

A third relevant question addresses whether inattentive investors have *delayed* reactions to narrative content. That is, when inattentive investors rely on heuristic algorithms, do they eventually reassess narrative content more thoroughly? Such behavior is predicted by a number of theoretical papers and implies that inattentive investors react to additional narrative attributes, but only with a lag (e.g., Hendershott et al. [2021]; Hirshleifer and Teoh [2003]; Peng and Xiong [2006]). Empirically, this suggests that as inattentive investors eventually become attentive, share prices of coverage-firms continue to adjust to additional dimensions of narrative content other than tone. If, however, inattentive investors make an initial assessment that they never update, we would not expect narrative content to influence stock prices in the post-announcement period.

Additionally, we predict that non-coverage firms will have smaller post-announcement reactions compared to coverage firms. This follows because investors in non-coverage firms are expected to spend more time independently interpreting narrative content on the 10-K filing date because they cannot effectively delegate information processing to analysts. This does not imply that all investors in non-coverage firms are attentive; inattentive investors are likely present in all companies. The key point is whether their influence is significant enough to generate pricing errors.

On net, we expect non-coverage firms to attract an investor clientele that is more engaged and pays more attention to contemporaneous signals. This implies that the marginal investor pays more attention to 10-K narrative content, which leads to more efficient stock price reactions. It then follows that there should be no post-filing date date reaction to narrative content, or to the extent that there is one, it will be significantly less than that observed for coverage firms.

We begin our analysis by exploring how the narrative contained in the Management, Discussion, and Analysis (MD&A) section of the 10-K report is used to provide *context* to quantitative results and then explore how investors *react* to it. We extend Li [2010] by using a Naïve Bayes (NB) classifier [Hastie et al., 2009] to identify a multi-dimensional set of linguistic attributes that include: tone, optimism, pessimism, specificity, vagueness, directness, evasiveness, passivity, and aggressiveness. An innovative aspect of our paper is that we develop our set of NB classifiers via a crowd-sourced online survey.

We characterize how context is provided by considering the association between financial performance and each of the above narrative attributes. We perform several tests after controlling for the endogenous selection of the defined narrative attributes. First, we examine the association between reported earnings and narrative attributes and find that firms with strong earnings performance tend to convey a positive and optimistic message that avoids evasive and aggressive content. Second, we examine investor reactions to different narrative dimensions. We formally address the idea that inattentive investors may rely on tone-based heuristics by separately evaluating sub-samples based on analyst coverage. Third, we address situations where tone (alone) may not provide sufficient guidance on content. To do this, we identify situations when tone and quantitative results are inconsistent (e.g., narrative tone is positive and the earnings surprise is negative).[6]

To improve identification, we exploit the fact that some firms stagger the reporting of financial performance and the filing of their 10-K reports. These firms typically release summary press releases and host conference calls to disclose key quantitative information (e.g., earnings, revenues, etc.) prior to the 10-K filing date. We define companies that follow staggered reporting strategies as *pre-announcers*. By conditioning on pre-announcers, we can more efficaciously isolate investor reactions to the incremental information contained in MD&A sections. Our empirical design does not completely isolate the impact of narrative MD&A content. For example, 10-K's include detailed quantitative information in footnotes [Amel-Zadeh and Faasse, 2016]. However, on balance, our focus on filing date returns for pre-announcers removes a significant amount of confounding information related to key performance metrics and significantly increases our power to test the incremental impact of MD&A narrative content.

In addition to applying a strategy to identify higher dimensional attributes, we address the countervailing issue of whether narrative text is swamped by boilerplate language. Part of our research design is devoted to the training of a specific NB classifier to separate voluntary disclosure from standard, often repetitive, legal and financial (i.e., regulatory driven) text which we refer to as boilerplate clauses.[7] This removes narrative content to which investors

[6]Remarkably, this comprises nearly 50% of the sample and so can be described as a relatively common occurrence. That is, we consider how results vary for a classic 2×2 stratification of the data (e.g., covered vs. not-covered by analysts with tone and surprise consistent or not consistent). Then, for each cell we assess whether there is any evidence that the additional attributes of narrative are associated with investor reactions.

[7]We argue that removing narrative content provided to satisfy mandatory regulatory statutes that require coverage of some topics in a largely prescribed and standardized fashion, allows us to better identify the voluntary narrative choices of management. This step is particularly relevant since, when the proportion of boilerplate disclosure in a given document is high, it raises a significant issue for the application of any Natural Language Processing (NLP) technique. If one wants to develop a model linking disclosure in filings, to some economically determined choice variable, such as returns, theory suggests that voluntary (non-boilerplate)

are unlikely to react or pay attention to. We find that the use of boilerplate text is common and comprises approximately one third of all narrative text under study.[8]

The rest of the paper is organized as follows. In section 2, we review the relevant literature. In section 3, we introduce and discuss our methodology for textual classification. In section 4 we formalize our set of hypotheses. In sections 5 and 6 we present and discuss the empirical specifications and related results, respectively. In section 7, we draw our conclusions.

2 Background

2.1 Investor Inattention

Limited investor attention impedes the process of incorporating public information into security prices. There is an extensive literature on investors' limited attention. A number of theoretical models have been developed that explain how investor inattention results in security prices that gradually adjust to public information (e.g., Hong and Stein [1999]; Hirshleifer and Teoh [2003]; Peng and Xiong [2006]; Andrei and Hasler [2015]; Hendershott et al. [2021]).

A growing body of empirical research suggests that investor inattention can lead to underreaction to value-relevant information. These studies show that this information comes from a variety of different sources, such as new products, earnings news, investor demographics (e.g., retail and institutional investors), analyst inattention, innovative efficiency, liquidity shocks, or information about related stocks (e.g., Huberman and Regev [2001]; Hirshleifer et al. [2004]; Hou and Moskowitz [2005]; Cohen and Lou [2012]; Hong et al. [2007]; DellaVigna and Pollett [2007]; Barber and Odean [2008]; Cohen and Frazzini [2008]; Hirshleifer et al. [2009]; Hirshleifer et al. [2013]; Bali et al. [2014]; Guo et al. [2020]; Cohen et al. [2020]). One of the themes that emerges from this literature is that investors process information that is easier to understand rapidly and postpone analysis of more complex less salient information.

Several papers use investor inattention to motivate the use of heuristics that are designed to simplify problem solving (e.g., Barberis and Shleifer [2003]; Hong and Stein [2007]; Peng and Xiong [2006]; Huang et al. [2014]). This literature contemplates the possibility that simple heuristics can be used to form an initial assessment of a financial statement that may be subsequently updated as investors have the time to delve more deeply into its content.

As 10-Ks have become longer and more complex, it takes more time for investors to process information. This leads to delayed prices reactions as investors gradually unpack the information contained in financial reports. Cohen et al. [2020] conclude that there is "an extreme

disclosures need to be identified [Dye, 1985].

[8] This is comparable to the proportion of boilerplate identified in the test sample. It's removal helps isolate the topics that regulators encourage firms to voluntarily discuss in their own words.

broad-based form of investor inattention to an item that is foundational to the corporate reporting process – the quarterly and annual reports – which leads to large return predictability." They find that firms that have significant changes in their 10-K and 10-Q reports outperform non-changers 34-58 basis points per month over the *following* year. They show that the reporting changes are concentrated in the MD&A section of the reports. Their findings are intuitively compelling because the MD&A secrion is the one place where the SEC encourages management to "tell their own story." Cohen et al. [2020] also note that "It is not merely the difference between quantitative and qualitative information that matters for investors (as in Engelberg [2008], but also the way in which that qualitative information is constructed and presented." Our study expands their analysis by evaluating whether investors respond (with a lag) to specific narrative aspects of MD&A.

2.2 Narrative Content

Significant portions of a firms' annual 10-K filings contain narrative that helps to contextualize financial statements and among these, the most important is the Management Discussions and Analysis (MD&A) section. MD&A is of particular interest because it is the one place in a 10-K where regulators actively encourage firms to "give the company's perspective on the business results of the past financial year. [...] The MD&A allows a company's management to tell its story in its own words." [SEC, 2020].

There is a related body of literature that focuses on financial narrative that is broad in scope and voluminous. A number of recent papers survey the relevant research in this area [Li, 2010; Loughran and McDonald, 2016; Lewis and Young, 2019; El-Haj et al., 2019]. Rather then repeating much of that discussion here, we focus on a relatively small set of papers that directly inform our analysis. We separately discuss papers that use three widely employed natural language processing techniques: context-specific dictionaries ([Loughran and McDonald, 2011]), Naïve Bayesian classification ([Hastie et al., 2009]), and topic modelling ([Blei et al., 2003]).[9]

Dictionary-based approaches estimate quantitative metrics based on counts of words that are included in content-specific dictionaries. Loughran and McDonald [2011] have designed a popular set of financial dictionaries that allow researchers to calculate positive and negative tone.[10] Examples of papers that use context-specific dictionaries include those that have been

[9]Other textual methods are frequently used in the literature either separately or in conjunction with the methods discussed here. These include readability scores (the FOG index), document length, key word searches (and associated word counts), and cosine similarity. These approaches are discussed in the survey papers mentioned above.

[10]Loughran and McDonald [2011] also provide dictionaries to evaluate additional attributes. These include: uncertainty, litigious, strong modal words, and weak modal words. The finance-specific dictionaries generally

calibrated to financial applications ([Davis and Tama-Sweet, 2012], [Feldman et al., 2010], [Loughran and McDonald, 2011]), while other studies employ generalist dictionaries ([Mayew et al., 2015] and [Muslu et al., 2014]).

Most of these studies use tone (or sentiment) as their main independent variable to evaluate future accounting performance (e.g., earnings) and stock market reactions. Results are mixed in the sense that tone seems to be a good explanatory variable in some cases, but not all. These studies share a common theme in the sense that they focus on either tone or the identification of specific phrases that reflect, for example, forward-looking statements.

One exception to the exclusive use of tone-based metrics is Muslu et al. [2014], which examines the association between *optimistic* forward-looking statements in Management, Discussion, and Analysis (MD&A) sections of 10-k filings and future earnings informational content. This study is particularly noteworthy because, to the best of our knowledge, it is the first to systematically examine a narrative attribute other than tone. The authors use a dictionary-based approach to detect evidence of optimism in narrative and conclude that it has incremental explanatory power relative to tone.

Naïve Bayesian Classification (NBC) is another class of natural language processing techniques that has been used to study narrative content. NBC is a supervised learning method that has not has gained as much traction among academics as dictionary-based approaches or topic modeling. This is likely attributable to the significant time and effort required to build classifiers. A key advantage of NBC relative to dictionaries is that it accommodates a more precise specification of the underlying research question. For example, Ryans [2019] uses NBC to classify firms that are associated with future write-down and restatements.[11]

Li [2010] is the first paper to use NBC. He evaluates the association between tone and various measures of financial performance and shows that tone is positively associated current performance, small size, future earnings, and low growth opportunities. Our paper builds upon Li [2010] by extending the set of language characteristics that are available for study. In Li's own words: *"[...]Conceptually, there are at least three disclosure characteristics that are interesting to researchers: the level (how much you say), the tone (what do you mean), and the transparency (how you say it).[...]"*. This is where we depart from his work. By expanding the number of linguistic attributes, we can more fully address "how you say it." This allows us to determine whether other dimensions of narrative content detected in MD&As partially explain reported earnings and subsequent investor reactions.

track the attributes characterized in generalist dictionaries. The difference is that words with ambiguous interpretations are deleted from the more generalist wordlists (e.g. Harvard General Inquirer [Stone et al., 1966]).

[11] A study by Henry and Leone [2016] compares dictionary-based approaches to NBC and concludes that dictionary-based approaches work as well as NBC in many but not all settings.

The third and final technique is topic modeling. Topic models assume that a document can be treated as a (probabilistic) mixture of topics themselves defined by a mixture of words. The most notable of these approaches is Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. One of the advantages of LDA is that it is an unsupervised method that does not rely on researcher pre-judgement. Many papers have successfully used LDA (Bao and Datta [2014], Dyer et al. [2017], Hoberg and Lewis [2017], Huang et al. [2018], and Brown et al. [2020]).[12] As was the case with our discussion of context-specific dictionaries, this approach and the relevant literature are discussed at length in the surveys mentioned above.

3 Research Design

This section presents the research design we adopt to study financial narrative and to test its explanatory power. We discuss our approach to identifying multi-dimensional language features and detail how we build the textual classifier to identify them. We finally discuss the characteristics of our training set and the predicted attributes.

3.1 Narrative as a Multi-Dimensional Feature Space: Giving Context to Financials

Human language is a complex construct. This implies that when studying a passage of text, individuals may have different perspectives that lead to different interpretations. In addition to the study of tone, emotion detection also is relevant and sets the stage for our methodology. Works such as Elliott [1992], Read [2004] and Neviarouskaya et al. [2009] are seminal contributors to research that focuses on the detection of multi-dimensional semantic features in narrative. This culminated with the emotion lexicon introduced by Mohammad and Turney [2010] who identified and explored several latent characteristics of text. We follow

[12]For instance, an early paper in the accounting literature by Bao and Datta [2014] used LDA to identify risk types from risk disclosures found in Item 1A of 10-Ks, and evaluate how the identified profiles affect the risk perceptions of investor. Dyer et al. [2017] identify specific trends in 10-K disclosure over the period 1996-2013 like the increase of length and boilerplate text with respect to the release of any related hard information (i.e., financial statement data). Through LDA, they capture focused topics that explain the increase in document length after the new FASB and SEC requirements. Hoberg and Lewis [2017] analyze 10-K MD&A sections and find that fraudulent managers tend to discuss fewer details around firm's performance but strategically talk more about positive aspects of the same firm's performance. Huang et al. [2018] study the analyst information intermediary roles in analyst reports and corporate disclosure. They find that analysts tend to discuss very specific topics that go beyond those covered in earnings conference calls but interpret the topics that were discussed in the conference call. Also, investors pay extra attention to new information in analyst reports when management has incentives to refuse to disclose value-relevant information in conference calls. Brown et al. [2020] use LDA to evaluate whether the thematic content in 10-Ks is incrementally effective and informative in detecting and predicting intentional misreporting. The identified topics together with the related attention on them are both effective signaling mechanism for a correct detection of financial misreporting.

the spirit of this approach but rely on machine learning, instead of lexicons, to capture the features of interest thus extending Li [2010].

The collection of MD&As under study is the starting point of our algorithm. More mathematically, we define a corpus C as a collection of N documents such that $C = \{D_i\}_{i=1}^{N}$. Any given document D_i can be represented by a $l-$dimensional space $\Omega_l(D_i)$ of language attributes. We assume that each dimension provides a different perspective and covers specific content around the narrative contained in document D_i. We also relax the assumption of independence of the attributes and allow for an explicit dependency structure across them. With this framework in mind, we can define a general equation describing the level of understanding of a document D_i as a function $f(\xi)$ of its content, parametrized by a set of features $\Omega_l(D_i)$, as follows:

$$U^l(D_i) = \int_{\Omega_l} f(\xi)\mathrm{d}\xi. \tag{1}$$

Solving Eq. (1) over the whole l-dimensional space would result in a complete and perfect understanding of the document D_i. We can decompose any document D_i in a $p-$dimensional sub-space $\Omega_p(D_i) \subseteq \Omega_l(D_i)$ defined by a set of attribution features (i.e., *attributes*) representing the different language characteristics we want to study.[13] This greatly simplifies Eq. (1) since the representation of D_i is now mapped by a lower dimensional space $\Omega_p(D_i)$ and is now discrete. We can then rewrite Eq. (1) as follows:

$$U^p(D_i) = \sum_{p=1}^{P} f(\xi)_p. \tag{2}$$

The theoretical framework described by Eq. (2) can be easily implemented on real data by empirically testing whether components other than tone could influence a specific outcome of interest. More practically, our study focuses on five attributes: tone (positive and negative), optimism (optimistic and pessimistic), specificity (specific and vague), directness (direct and evasive) and aggressiveness (aggressive and passive). The sum in Eq. (2) is then truncated at $p = 5$ and each given sentence in our corpus is represented as a combination of $\Omega^*_{p=5}$ sub-space with each component set on a scale $\{-2, \ldots, +2\} \in \mathbb{Z}$. The components with the related scores are described in Table 1.[14]

[13] Asymptotically, the full mapping of the feature space occurs when $p \to l$. For both practical and computational limitations, we usually have $p \ll l$. Moreover, when $p \to l$ we are forced to work with extremely sparse matrices which can make the inference intractable. Notwithstanding, this does not change our intuition and could only affect the level of understanding of the narrative content as in any other dimensionality reduction problem.

[14] We deliberately drop the dependency on D_i for clarity purposes.

Table 1: The five proposed attributes with their scores and respective levels as defined in the training set. SW. stands for *Somewhat*, for brevity.

Scale	**Tone**	**Optimism**	**Specificity**	**Directness**	**Aggresiveness**
-2	Negative	Pessimistic	Vague	Evasive	Passive
-1	SW. Negative	SW. Pessimistic	SW. Vague	SW. Evasive	SW. Passive
0	Neutral	Neutral	Neutral	Neutral	Neutral
+1	SW. Positive	SW. Optimistic	SW. Specific	SW. Direct	SW. Aggressive
+2	Positive	Optimistic	Specific	Direct	Aggressive

We support our intuition with the following examples taken from different MD&As.

> ***Sentence A***. We expect production for this program to ramp up over the next several years.
>
> ***Sentence B***. Since our initial registration we have: recycled over 800,000 pounds of paper, cardboard and plastic; reduced electricity usage by an average of 5% per year; and reduced natural gas usage by an average of 3% per year.

The above two sentences share a positive tone, but have fundamentally different content. *Sentence A* is a forward-looking statement while *Sentence B* is a statement of a fact. For instance, one can argue that *Sentence A* is positive and shows a somewhat bold and aggressive spin. Conversely, *Sentence B* is also positive, but is purely factual and does not show any element of spin. Building on this assumption, one can investigate more complex structures within the same construct. For example, one might notice how there is a higher degree of specificity in *Sentence B* due to the details that are provided through quantitative numbers. Conversely, we do not have any specific information in *Sentence A* which, at this point, looks like a forward-looking statement without substantive support.

To further substantiate our framework of analysis, we show two sentences with negative tone.

> ***Sentence C***. Due to the fact that we have not generated any revenues, we believe that the financial information contained in this Annual Report on Form 10-K is not indicative of, or comparable to, the financial profile that we expect to have once we begin to generate revenues.
>
> ***Sentence D***. We have experienced significant fluctuations in quarterly shipments and revenues and, beginning in the fourth quarter of 2008, we saw many potential customers lengthen their sales cycles and postpone purchase decisions.

These examples clearly illustrate how tone and optimism are not necessarily correlated. We can observe how *Sentence C* states a negative fact related to the firm not generating any revenue, although with an optimistic forward-looking statement about future revenue generation. Conversely, *Sentence D* conveys a negative tone together with a pessimistic forward-looking statement related to customers postponing purchases.

It is more evident now how we can decompose a text (e.g., a sentence) into sub-components. Each component contributes to the overall understanding of the passage according to Eq. (2). We believe that focusing on just one attribute (e.g., tone) to characterize the narrative is not only insufficient but also introduces limitations by construction. In Table 2, we illustrate how the two coupled representative sentences, which share an identical tone score, differ with respect to all other components. This illustrates and clarifies our main conjecture that identical tonal scores do not necessarily imply the same overall content. In this paper, we differentiate and disentangle this issue by independently assessing each attribution feature's influence.

Table 2: Predicted attribute scores for *Sentence A.*, *B.*, *C.*, and *D.*. SW. stands for Somewhat, for brevity.

Sentence	**Tone**	**Optimism**	**Specificity**	**Directness**	**Aggressiveness**
Sentence A.	Positive	SW. Optimistic	SW. Vague	SW. Evasive	Aggressive
Sentence B.	Positive	Neutral	Specific	Direct	Neutral
Sentence C.	Negative	Optimistic	Specific	Direct	SW. Aggressive
Sentence D.	Negative	Pessimistic	SW. Specific	SW. Direct	Neutral

3.2 Supervised Detection of Attribute Features

To assess the narrative content of MD&A sections, we construct Naïve Bayes classifiers [Hastie et al., 2009] for tone, optimism, specificity, directness, aggressiveness, and boilerplate using a web-based survey tool. Responses are crowd-sourced from a pool of 283 respondents that include members of the authors LinkedIn networks and students enrolled in graduate accounting and finance classes. Details of the construction of the Naïve Bayes classifiers are contained in Appendix A.[15]

Crowd sourcing the data collection phase of our project has several advantages. First, it ensures that the results are not subject to personal biases of a small set of individuals. Second, the survey is designed to minimize the risk of "classification fatigue" by allowing respondents

[15] We were able to survey individuals both in academia and with a wide variety of professional backgrounds and expertise in financial services, auditing, consulting, asset management and financial regulation. Although it would be interesting to analyze classification differences across respondents, we did not collect demographic data because we believe that anonymity maximizes response rates. The survey is designed to remain open indefinitely and the number of classified sentences grows over time.

to score as many sentences as they want. In addition, each respondent can connect to the survey tool with no limits nor restrictions in terms of number of accesses or duration of the session. These features increase our ability to collect a large amount of data in a timely and accurate manner.

3.3 Data

Our sample period of individual firms' 10-K filings extends from 2001 through 2018. MD&A section (Items 7 and 7A) is extracted from 10-K reports downloaded as *Stage One Parse Data* from the Software Repository for Accounting and Finance of the University of Notre Dame [2021] [Loughran and McDonald, 2016]. These files have been already cleaned from extraneous texts such as HTML, ASCII-encoded segments, and tables.[16] To be included in the final sample, we require firms to have financial statement data in Compustat and returns data in the Center for Research in Security Prices (CRSP) database. Consensus earnings forecasts, realized earnings, and earnings report dates are collected from IBES. If a firm is not included in IBES, we augment the missing data with data from the Compustat Annual and Quarterly files. Cumulative abnormal returns are calculated using the Market Model. The market proxy is the CRSP Value-Weighted Index. We report descriptive statistics for the variables used in the paper in Table 3.

4 Hypothesis Development

We are interested in examining investors' responses to the incremental information contained in the MD&A section of 10-K reports. Firms prepare the narrative component of their 10-K reports with the full knowledge of the year's financial results. This creates an endogenous link between the narrative and quantitative components. To evaluate the incremental information in the MD&A, one needs to control for investor reactions to quantitative financial (revenues, earnings, etc.).

There is significant cross-sectional heterogeneity in the approaches firms employ to disclose financial results. Many choose to hold earnings conference calls on the same day they file their 10-K reports with the SEC. Some of these firms may simultaneously issue a press release, which, if material, would require the filing of an 8-K. Other firms may instead forgo an earnings call and simply issue a press release, particularly those that do not have analyst coverage. Still, others follow a staggered reporting strategy in which they issue a press release, possibly in

[16]For a thorough description of the pre-processing steps and parsing details, we refer the reader to the official repository page at `https://sraf.nd.edu/data/stage-one-10-x-parse-data/`. The filings can also be downloaded, for free, from the same page.

conjunction with an earnings conference call, *before* they file their 10-K report. Alternatively, they may host a pre-filing earnings call without issuing a formal press release.

Our empirical design explicitly recognizes the difficulty in disentangling the incremental effects of narrative content from quantitative numbers, especially for firms that simultaneously disclose both. Since voluntary narrative text is provided to give context to the quantitative results, quantitative and qualitative content are likely to be positively correlated. This results in a potential multi-collinearity problem that makes it challenging to detect and isolate significant effects uniquely associated with narrative content. We address this concern by focusing on *filing date* reactions for firms that follow *staggered* reporting strategies.[17] Our focus on firms that have pre-announced earnings ameliorates some of the concerns about the potential conflation of narrative content with quantitative metrics.

Our first hypothesis predicts that narrative content is influenced by strategic selection of narrative attributes. Earlier research has demonstrated a positive association between earnings and tone. We are interested in expanding this inquiry to include additional linguistic attributes (Optimism, Specificity, Directness, and Aggressiveness) and to incorporate additional features of firms' operating environment beside earnings. We characterize the operating environment using a set of control variables that reflect financial performance (Earnings, Free Cash Flow, Earnings Volatility), operational characteristics (Number of Business Segments, Herfindahl Concentration Index, Growth Opportunities), changes in operating environment (Recent M&A Activity, Downsizing of Operations), and financial condition (Leverage).[18] We formalize our first hypothesis as follows:

Hypothesis 1 (H1): *Firms use multi-dimensional linguistic attributes to give context to reported financial performance and that the strategic choice of narrative content is influenced by firms' operating environments.*

We extend the univariate analysis proposed in Hypothesis 1 by evaluating the extent to which firms emphasize multi-dimensional narrative attributes to achieve an *overall* disclosure objective. To do this, we estimate an Aggregate Attribute Index (AAI) that projects individual attributes into a a single metric. This allows us to consider whether an aggregate measure of narrative content is associated with features of firms' operating environments.

[17]While firms disclose key metrics that are important to investors on the earnings announcement date, a 10-K will contain residual quantitative information not previously disclosed. We consider this information to be of second-order importance, particularly as it relates to the information disclosed in earnings conference calls where analysts are permitted to ask questions about the issues they deem to be the most value-relevant. Nevertheless, we believe this information still provides additional insights which attract investors attention.

[18]We use the Book to Market ratio to capture growth opportunities.

Hypothesis 2 (H2): *Firms emphasize individual narrative attributes to different degrees. Aggregate narrative disclosure strategies provide context to financial performance and are predicted to be associated with the operating environment.*

We propose that the desire for firms to provide higher dimensional narrative context depends upon the overall information environment. Specifically we hypothesize that firms with analyst coverage will provide narrative content that more closely tracks financial performance. This implicitly assumes that monitoring by analysts makes firms less likely to deviate from "script." Such behavior does not necessarily imply that narrative content is more informative. It is only meant to suggest that content will be more correlated with performance. By contrast, when firms do not have analyst coverage, they may use the MD&A to provide more value-relevant content in the MD&A because investors do not have access to analyst research. This could, for example, result in non-coverage firms providing more forward-looking information (e.g., optimism). This is relevant here because forward-looking content, like optimism, would be less likely to track current accounting performance. For example, a firm with poor earnings performance may try to reassure investors that the current year is an aberration by including optimistic content. Alternatively, non-coverage firms may intentionally choose to provide less narrative context to create opacity. We summarize these ideas in Hypothesis 3:

Hypothesis 3 (H3): *Firms will be more likely to provide multi-dimensional narrative content that is associated with financial performance if they have analyst coverage.*

Hypotheses **H1** through **H3** make predictions about how firms craft corporate communications to help explain financial performance. An equally interesting line of inquiry is to consider how investors react to narrative content. We make specific cross-sectional predictions that follow from the investor inattention hypothesis.

If inattentive investors adopt simple heuristics to evaluate narrative content, it may result in incomplete reactions to the information contained in 10-K filings that result in pricing errors [Hendershott et al., 2021]. We further hypothesize that the opportunity to free-ride on analyst research allows inattentive investors to substitute their normal due diligence with tone-based heuristics. By contrast, investors in non-coverage firms must perform their own due diligence, implying that, on average, they are more attentive.

Hypothesis 4 (H4): *On the 10-K filing date, investors in firms that pre-announce earnings and have analyst coverage are more likely to react to the tone of an MD&A and place less emphasis on other narrative attributes.*

Hypothesis 5 (H5): *On the 10-K filing date, investors in non-coverage firms that pre-announce earnings tend to react to multi-dimensional narrative attributes (tone, optimism, etc.).*

A key prediction of the investor inattention hypothesis is that there will be a delayed reaction to narrative content. In other words, inattentive investors eventually become attentive and, to the extent that MD&A contains value-relevant information, they are expected to react to multi-dimensional narrative content.

If investors initially use tone-based heuristics to make initial assessments on the filing date, the willingness to subsequently evaluate the complete narrative may depend on the consistency between narrative tone and earnings surprises. We hypothesize that inattentive investors will pay more attention to the multi-dimensional aspects of the MD&A when tone and earnings surprises are inconsistent. For example, a firm that reports unexpectedly poor earnings but tries to put a positive spin on the reported performance is more likely to cause a previously inattentive investor to dig into the narrative content.

Hypothesis 6 (H6): *Relative to non-coverage firms, investors in firms that pre-announce earnings and have analyst coverage have a greater tendency to evaluate the full narrative content of MD&A with a lag. This effect is expected to be more pronounced in cases where tone and earnings surprises are inconsistent.*

5 Empirical Specifications

We are interested in examining how firms structure the narrative content of MD&A sections and how investors respond. A key factor motivating our analysis is that managers have direct and private access to the information they are about to make public before they draft the MD&A section. This provides management with an opportunity to adopt reporting strategies designed to influence investor expectations. At issue is whether managers simply focus on tone or strategically employ other narrative attributes.

5.1 Association Between Narrative Attributes and Reported Earnings

Hypothesis H1 considers whether there is an association between narrative content and reported earnings (*Earnings*). The test relies on the following regression specification:

$$Attribute_{i,t} = \alpha + \beta \cdot Earnings_{i,t} + \boldsymbol{\gamma} \cdot \mathbf{controls}_{i,t} + FE_{Firm} + FE_{Year} + \epsilon_{i,t}, \tag{3}$$

where $Earnings_{i,t}$ is the percentile of Net Income from year t scaled by Total Assets from year $t-1$. $Attribute_{i,t}$ denotes the Naïve Bayes classification scores for *Tone*, *Optimism*, *Specificity*, *Directness*, and the absolute value of *Aggressiveness*. The vector of firm-specific control variables is denoted as $\mathbf{controls}_{i,t}$, FE_{Firm} denotes firm fixed effects, and

FE_{Year} denotes year fixed effects. We use percentiles instead of levels because the attribute factors are integer-valued.

An intrinsic limitation of this specification lies in its inability to consider cases in which attributes are influenced by latent factors that managers take into account when drafting the financial statements. To address this, we conduct a second test of hypothesis **H1** that reverses Eq.(3) and treats $Earnings_{i,t}$ as the dependent variable. An advantage of this specification is that we can directly control for the endogenous selection of narrative content using a Heckman-type adjustment [Heckman, 1979]:

$$\begin{aligned} Earnings_{i,t} &= \alpha + \beta_1 \cdot Tone_{i,t} + \beta_2 \cdot Optimism_{i,t} + \beta_3 \cdot Specificity_{i,t} \\ &+ \beta_4 \cdot Directness_{i,t} + \beta_5 \cdot |Aggressiveness_{i,t}| + \boldsymbol{\gamma} \cdot \mathbf{controls}_{i,t} \\ &+ \eta \cdot \lambda_{i,t} + FE_{Firm} + FE_{Year} + \epsilon_{i,t}, \end{aligned} \tag{4}$$

where $\lambda_{i,t}$ is the inverse Mills ratio.

5.1.1 Endogenous Selection of Narrative Content

To control for narrative attribute selection, we define an Aggregate Attribute Index ($AAI_{i,t}$). The construction of $AAI_{i,t}$ is straightforward and follows the approach used by Gompers et al. [2003] to calculate their Governance Index. We define $AAI_{i,t}$ as a linear combination of the Naïve Bayes classification scores over all five attribute classification levels:

$$AAI_{i,t} = Tone_{i,t} + Optimism_{i,t} + Specificity_{i,t} + Directness_{i,t} - |Aggressiveness_{i,t}|. \tag{5}$$

For example, $Tone_{i,t}$ denotes the estimated classification level for the Tone attribute that is associated with the MD&A for firm i at time t. We subtract the absolute value of $Aggressiveness_{i,t}$ because both aggressiveness and passivity have negative connotations. This treats passive and aggressive scores symmetrically.

We truncate $AAI_{i,t}$ at (-4) and (+4) because the number of observations that fall outside this range is small.[19] Although giving each attribute equal weight does not consider their

[19] We perform robustness tests to determine if our results are sensitive to truncation. First, we estimate $AAI_{i,t}$ using the first principal component of the attribute classification levels. Second, we estimate $AAI_{i,t}$ as the sum of the attribute classifications that have the highest cosine similarity with each attribute's Naïve Bayes classifiers. As an example, the most likely classification score for firm i's *Tone* at time t is estimated by determining the maximum cosine similarity of sentence j's word proportion vector with the word proportion vector of each of its *Tone* attribute levels. Since the results are qualitatively similar for both variables specifications, we conclude that our proposed method is robust to truncation. That is, it is not sensitive neither to

relative importance, Gompers et al. [2003] note that this type of index has the advantage of being transparent and easily reproducible. Using $AAI_{i,t}$ as a dependent variable, we then estimate a random-effects ordered probit model [Crouchley, 1995] as follows:

$$AAI_{i,t} = \beta \cdot x_{i,t} + \nu_i + \epsilon_{i,t}, \tag{6}$$

where $x_{i,t}$ is a vector of independent covariates, ν_i is the panel random effect for firm i at time t, and $\epsilon_{i,t}$ is the stochastic disturbance term. Both ν_i and $\epsilon_{i,t}$ are assumed to be independent and identically distributed as $\mathcal{N}(0, \sigma_\nu^2)$ and $\mathcal{N}(0, \sigma_\epsilon^2)$, respectively.

According to this model, the probability of observing firm i's response at time t is uniquely determined as:

$$Pr\left(AAI_{i,t} > k | \kappa, x_{i,t}, \nu_i\right) = \Phi\left(\beta \cdot x_{i,t} + \nu_i - \kappa_k\right), \tag{7}$$

where k denotes the cardinal values of the aggregate attribute score ($k = \{-4, \ldots, 4\}$), κ_k denotes the cutpoint k, and $\Phi(x_{i,t})$ is the standard normal cumulative distribution function. The cutpoints κ_k are used to estimate the probability of observing an attribute score equal to k such that:

$$Pr\left(AAI_{i,t} = k | \kappa, x_{i,t}, \nu_i\right) = \Phi\left(\kappa_k - \beta \cdot x_{i,t} - \nu_i\right) - \Phi\left(\kappa_{k-1} - \beta \cdot x_{i,t} - \nu_i\right), \tag{8}$$

Because $AAI_{i,t}$ is not observed, its location is fixed by setting the intercept in Eq.(3) equal to zero. Lastly, the inverse Mills ratio for the ordered probit model of Eq.(6) is calculated as follows:

$$\lambda_{i,t} = \frac{\phi\left(\hat{\kappa}_j - \beta \cdot x_{i,t} - \nu_i\right) - \phi\left(\hat{\kappa}_{j+1} - \beta \cdot x_{i,t} - \nu_i\right)}{\Phi\left(\hat{\kappa}_j - \beta \cdot x_{i,t} - \nu_i\right) - \Phi\left(\hat{\kappa}_{j+1} - \beta \cdot x_{i,t} - \nu_i\right)}, \tag{9}$$

where $\hat{\kappa}_j$ is the estimated cutpoint for classification level j.

5.2 Summary of Model of Investor Reactions to Narrative Content

Our primary hypotheses evaluate how investors react to the narrative content of MD&A disclosures. We test our primary hypotheses using a regression model that treats cumulative abnormal returns (CARs) as the dependent variable. The exogenous variables include narrative attribute metrics, firm-specific control variables ($\mathbf{controls}_{i,t}$), a Heckman-correction to control for endogenous attribute selection ($\lambda_{i,t}$), and year fixed effects (FE_{Year}).[20] We

the truncation level of choice nor to the procedure used to estimate attribute classification levels.

[20] We do not include firm fixed effects because CAR is expected to have mean zero. We include year fixed effects because CARs may be correlated within a given year as we expect to observe positive cross-

estimate the following model:

$$CAR_{i,t} \quad = \quad \alpha + \beta \cdot \mathbf{Attribute}^{M}_{i,t} + \boldsymbol{\gamma} \cdot \mathbf{controls}_{i,t} + \eta \cdot \lambda_{i,t} + FE_{Year} + \epsilon_{i,t}, \qquad (10)$$

where $\mathbf{Attribute}^{M}_{i,t}$ denotes a vector of attribute metrics.

We evaluate three different specifications: (1) Investor reactions on the 10-K filing date for coverage and non-coverage firms. This specification is used to test Hypotheses 4 and 5. (2) Investor reactions to narrative attributes on the 10-K filing for coverage and non-coverage firms that are further stratified by the consistency between earnings surprises and tone. (3) Investor reactions to narrative attributes in the 60-day trading period following the 10-K filing date for coverage and non-coverage firms that are further stratified by the consistency between earnings surprises and tone.

6 Empirical Results

6.1 The Determinants of Multi-Dimensional Narrative

This section examines the the extent to which firms align their MD&A disclosures with verifiable, firm-specific characteristics. We classify firm characteristics into four categories: reported firm performance, the operating environment, changes in the operating environment, and financial structure.

Our narrative attribute metrics are designed so that positive values are associated with favorable narrative content (positive, optimistic, specific, and direct) and negative values have unfavorable connotations (negative, pessimistic, vague, evasive, aggressive, passive). Taken as a whole, Table 4 provides evidence that is broadly consistent with predictions related to univariate attributes (Hypothesis **H1**) and the overarching disclosure strategy (Hypothesis **H2**).

This subsection is a straightforward examination of association. We estimate a series of regression models that treat individual narrative attributes as well as the AAI index as dependent variables. Given that inattentive investors are expected to use tone-based heuristics, Tone is the most important narrative attribute. To simplify the exposition, we refer to Optimism, Specificity Directness, and Aggressiveness as "higher-order" factors because they are primarily designed to provide additional context that helps clarify tone messaging. This classification scheme does not imply that higher-order factors cannot have independent value.

sectional correlation in standardized unexpected earnings (SUE), which would result in positive cross-sectional correlation in CAR.

6.1.1 Tone

Since Tone has both contemporaneous and forward-looking aspects, we expect Tone to be associated with measures of reported performance. The results in Table 4 indicate that firms design their MD&As in a manner that is strongly associated with the underlying economic determinants. We find that profitable firms with strong cash flows and relatively low operational uncertainty (earnings volatility) provide positive messages. Firms also tend to provide relatively upbeat narrative when they have relatively high growth opportunities (book-to-market) and have the debt capacity (leverage) to take advantage of them without facing significant capital constraints.

Table 4 indicates that when firms downsize operations they tend to be more negative. This suggests that firms do not attempt to put a favorable spin on an event that is typically viewed by investors as a sign of weak financial performance. We also find that acquisitive firms tend to provide direct content.

6.1.2 Optimism

Optimism is a higher-order classification of Tone because optimistic statements are typically positive in nature. However, optimistic statements are inherently forward-looking, while tone can convey contemporaneous and even backward-looking assessments of financial performance. Similar to our results for Tone, Table 4 indicates that high growth firms with relative low operating uncertainty tend to make optimistic statements. In addition, firms that downsize operations tend to be more pessimistic.

6.1.3 Specificity and Directness

Firms tend to incorporate specificity and directness when designing MD&As. As shown in Appendix A, Table A.1, over 95% of the MD&As are classified as at least somewhat direct and specific. There is, however, sufficient cross-sectional variation to conclude that high growth firms with low operational uncertainty tend to provide specific and direct narrative. Firms with high leverage also tend to provide specific and direct content. This suggests that firms try to provide clear and forthright narrative when their valuations are driven by future investment opportunities or they feel compelled to provide accurate assessments to creditors.

6.1.4 Aggressiveness

The Aggressiveness classifier assigns positive values to passive narrative and negative values to aggressive narrative. Since both attributes have unfavorable connotations, the dependent variable is calculated as the absolute value of Aggressiveness multiplied by -1. This transformation reorders the Aggregate classifications so that a negative value has a negative connotation. Table 4 reports that firms tend to be *less* aggressive when profitability is low and operating uncertainty is high. Complex firms with that operate in concentrated industries also tend to be less aggressive. Finally, highly levered firms tend to be more aggressive, but back these assertions up with more specific and direct content.

6.1.5 Multi-dimensional narrative content

The idea that firms provide additional context by intentionally combining Tone with higher-order narrative content is explored in this section. When viewed collectively, higher values of AAI indicate that the MD&A is more favorable.

As a first step, we examine the simple association between AAI and economic determinants using ordinary least squares regression. Table 4 finds that our primary univariate results continue to hold. That is, high growth firms (book-to-market) that are profitable (earnings, free cash flows) and have relatively simple operations (segments) with low levels of operating uncertainty (earnings volatility) tend to design favorable MD&As.

In our subsequent tests, we estimate a "Narrartive Choice" model that controls for the endogenous selection of narrative content by including a Heckman-style adjustment. Specifically, we estimate the inverse Mills ratio using a random-effects ordered probit model. Although the model is primarlily designed to address enogeneity, it is of independent interest because it provides additional insights into overall narrative design.

Table 5 reports the results of a random-effects ordered probit regression that treats the Aggregate Attribution Index (AAI) as the independent variables and uses the same independent variables as Table 4. Similar to the results in Table 4, we find that firms with high growth opportunities, relatively strong free cash flows, and low earnings volatility tend to provide more favorable narrative. In addition, acquisitive firms with simple operating structures also tend to have favorable MD&As.

6.2 Reported Earnings, Narrative Content, and Analyst Coverage

Equity research analysts help intermediate the flow of information between management and shareholders. Examples of this activity include direct access to management during the Q&A portion of earnings calls and bank-sponsored investor conferences. The combination of direct access and the production of research reports, forecasts of key performance metrics (earnings), and investment recommendations effectively makes analysts indirect disclosure monitors.

This raises the possibility that firms design their disclosure policies with different objectives, especially when they are subjected to greater external scrutiny. Hypothesis **H3** predicts that firms with analyst coverage will provide narrative content that tracks reporting earnings more closely than non-coverage firms. The underlying idea is that coverage firms will be reluctant to provide novel content that may cause analysts to reassess information that was previously provided during an earnings call. As a consequence, managers of coverage firms will prefer objective content and avoid subjective narrative that has not already been publicly discussed. By contrast, non-coverage firms do not have access to as many communication channels and are more inclined to include subjective, possibly forward looking content.

We test Hypothesis **H3** by regressing reported earnings (percentiles) on individual narrative attributes for subsamples of firms with and without analyst coverage. The results are broadly consistent with our predictions. We show that firms with analyst coverage report earnings that are significantly and "favorably" related to Tone, Optimism, Directness and Aggressiveness. By contrast, non-coverage firms report earnings that only are significantly and favorably related to Tone and Optimism. With the exception of Optimism, higher-order attributes (Directness, Specificity, and Aggressiveness) are not used as much, possibly because they are discussing important information that is not directly related to earnings.

Consistent with **H3**, Table 6 shows that the coefficient estimates for coverage firms are more strongly correlated with individual narrative attributes than those for non-coverage firms. For example, a firm with analyst coverage would report earnings that are almost one percentile higher ($2 \times (1.726 - 1.259)$) than non-coverage firms. If we evaluate all of the narrative attributes using their most favorable values, coverage firms would report earnings that are three percentiles higher than non-coverage firms.[21] We also note that the adjusted R-squared of the coverage firm model is 38% higher ((0.279-0.316)/0.202)) than the non-coverage firm model.

The results also indicate that the proportion of boilerplate narrative (*Boilerplate*) is insignificantly associated with *Earnings*, and that firms tend to include more sentences that are less verbose (lengthy). This suggests that boilerplate is included in MD&A to satisfy possible regulatory mandates but is not being used strategically by management to obscure the interpretation of financial performance. We also find that firms with relatively high earnings operate in industries where the proportion of firms that actively provide earnings guidance is relatively high. The coefficient on the Inverse Mills Ratio is statistically significant, indicating that it is important to control for narrative choice when evaluating its independent association

[21] Using the coefficients from Table 6, the difference in the earnings percentile for a coverage firm relative to a non-coverage firms is 3.014 ($2 \times (1.726 - 1.259) + 2 \times (1.556 - 1.353) + 2 \times (-0.343 + 0.064) + 2 \times (1.045 + 0.071) + 0 \times (0.880 - 0.316)$).

with reported earnings.

6.3 Investor Reactions to Narrative Content

This section tests the inattentive investor hypothesis with a series of abnormal return regressions that assess investor reactions on the 10-K *filing* date and in the 60-trading period following the filing date. Collectively, Hypotheses **H4** through **H6** establish a set of predictions that when taken together would be consistent with the presence of inattentive investors. Hypotheses **H4** and **H5** are designed to consider whether inattentive investor incentives to free-ride on analysts result in weaker reactions to narrative content on the filing date. According to Hendershott et al. [2021], inattentive investors react to important information with a lag, causing subsequent stock price reactions. This is formalized in Hypothesis **H6**, which also predicts that the reactions will be more pronounced if the earnings surprise and tone are inconsistent. That is, investors will eventually spend more time performing independent research when their initial assessment of tone on the *filing* date is at odds with the incremental information released on the *earnings announcement* date.

6.3.1 Filing Announcement Date Results

To test Hypotheses **H4** and **H5**, we estimate separate regressions for coverage and non-coverage firms. The dependent variable is the three-day cumulative abnormal returns (-1,+1) and the independent variables are narrative attributes and a set of control variables. For purposes of our tests, we decompose Aggressiveness into Passive and Aggressive components because investors may react differently to passive and aggressive content. To maintain consistency with our earlier analysis, we multiple Passive and Somewhat Passive values by -1, This implies that Passive and Aggressive attributes take values that range from 0 (a neutral response) to -2 (the most unfavorable response).

Table 7 reports results that are consistent with **H4** and **H5**. After controlling for firms' choice of narrative content, investors in firms with analyst coverage react positively to tone but do not attach significance to higher-order attributes. This is consistent with the hypothesis that inattentive investors use tone-based heuristics in conjunction with a tendency to free-ride on analysts.

One question that naturally arises is how do investors free-ride on analysts when they may not have time to update research reports or earnings forecasts? The simple explanation is that investors in firms that follow staggered reporting strategies have access to earnings conference calls that precede the 10-K filing date. On these calls, management would be expected to discuss results in a manner that is consistent with the narrative that will be disclosed when the 10-K is filed. Just as importantly, analysts have direct access to management during the Q&A portion of a conference call and may be able to extract valuable information, including forward-looking content. To the extent that this information is already publicly available when the 10-K is filed, investors will be less inclined to spend significant time processing the MD&A on the filing date.

By contrast, Table 7 indicates that, in addition to tone, investors in firms that do not have analyst coverage also respond positively to optimistic content. This suggests that investors use the narrative content of MD&A to make inferences about future cash flows that have not been previously identified by market participants. This finding is broadly consistent with Muslu et al. [2014] who find that investors in "dark" (no analyst coverage and no management forecasts) firms also respond positively to optimistic content.

6.3.2 Post-Filing Date Results

Table 8 reports investor responses over the 60 trading day window that follows the 10-K filing date. Panel A reports results for firms that provide consistent tone and earnings surprise messaging, and Panel B reports results for inconsistent messaging. Consistent with the inattentive investor hypothesis (**H6**), we find that investors in firms with analyst coverage have delayed reactions to higher-order narrative content regardless of whether tone and the earnings surprise are consistent. The regressions indicate that investors react to forward-looking (optimistic) narrative in MD&As only after they have had additional time to process it. That is, given the complexity associated with higher order narrative and the opportunity to free ride on analyst research efforts, investors are willing to defer a full evaluation of MD&A content until later. These findings are consistent with the literature on investor inattention [Hendershott et al., 2021] wherein it is shown that inattentive investors are slow to react to important information, resulting in mispriced securities.

While our post-filing date findings are consistent with that theory, our research goes further by suggesting that one potential reason for the slow reaction by some investors is that they defer making their own detailed assessments and rely instead on analyst research coupled with tone-based heuristics. Put simply, narrative takes longer to process than quantitative disclosures. This implies, for example, that in the absence of incentives to process complex narrative immediately, investors delay performing independent analysis that may be able to extract additional meaning from higher dimensions of narrative over and above tone.

We also find that investors in non-coverage firms fully react to narrative content on the filing announcement date. Consistent with Hypothesis **H6**, Table 8 indicates that none of the higher-order narrative attributes are significantly associated with post-announcement abnormal returns. This suggests that, as predicted, investors know that they must perform their own due diligence and are incentivized to do so on the filing announcement date. While it may still be the case that some investors process MD&As with a lag they are not the marginal investor in these stocks.

Hypothesis **H6** also predicts that investors in firms with analyst coverage are more likely to respond more aggressively to inconsistent narrative content. The idea is that inconsistent messaging may create enough confusion on the announcement date that investors will expend greater efforts in the post-announcement period to analyze the narrative. At first glance, the results do not appear to be support this prediction - optimism and to a lesser extent passivity are significantly associated with post-announcement abnormal returns regardless of message consistency. However, we note that the coefficient estimates do support this prediction in the sense that the coefficient estimate for Optimism in the Inconsistent message sub-sample (1.272) is 34.6% higher than the coefficient (0.945) in the Consistent messaging sub-sample.

Collectively, the results we present in Sections 6.3.1 and 6.3.2 support the inattentive investor hypothesis as characterized by Hypotheses **H4** through **H6**.

7 Conclusion

This paper tests and finds support for the inattentive investor hypothesis. The testing approach is based on the idea that 10-K filings contain complex narrative that requires considerable effort to process, This creates incentives for inattentive investors to use heuristic algorithms that allow them to rapidly process narrative. Given the relative ease of processing tone, we conjecture that inattentive investors use tone-based heuristics to help them process complex discussions in the MD&A section. The incentives to rely on heuristic algorithms is exacerbated when investors have the opportunity to free-ride in information produced by equity research analysts. We also hypothesize that when opportunities to rely on analyst research are unavailable, investor are more inclined to exert the effort required to process narrative content on the 10-K filing date.

We first evaluate whether firms strategically design narrative content in a way that accurately reflects underlying economic factors. We find that firm-specific factors which are generally considered important to investors, influence firms' narrative choices in intuitively obvious ways. While such a finding is reassuring, it demonstrates that firms link their MD&As to contemporaneous metrics.

Using a novel identification strategy that focuses on firms that follow "staggered" reporting strategies (i.e., firms that disclose financial results before filing their 10-Ks), we are able to evaluate the incremental information contained in the MD&A section. The benefit of this econometric design is that we mitigate concerns about the conflation of narrative with important quantitative performance metrics. We also control for firms' strategic design choices when crafting their MD&As by estimating a random-effects ordered probit model, which we use to control for possible narrative selection bias. Consistent with our version of the Investor Inattention Hypothesis, our results indicate that investors *only* react to tone messaging on the filling announcement date for firms with analyst coverage. This suggests that inattentive investors use tone-based heuristics in conjunction with analyst research to make initial assessments of narrative content. By contrast, investors in non-coverage firms react to tone **and** optimism on the filing announcement date indicating that these investors work harder to process information when they cannot free-ride on analyst research.

Support for transitory investor inattention is found in our post-announcement analysis. If investors are initially inattentive, they will process information with a lag. To the extent that MD&As provide valuable information, we expect to find delayed reactions to higher-order narrative attributes for firms with analyst coverage and that the reaction is expected to be stronger for firms that initially provide an inconsistent tone message (i.e., narrative tone is inconsistent with the earnings surprise). By contrast, we do not expect to find delayed responses for non-coverage firms because investors have incentives to fully process narrative content in the filing date. Our empirical results support this hypothesis.

Taken together, we provide evidence that is consistent with the prediction that inattentive investors delay processing complex narrative content when they have opportunities to free-ride on analyst research. One caveat is that while our findings are broadly consistent with investor inattention, we do not provide an unambiguous test of this hypothesis. For example, we make the intuitively plausible conjecture that investors use tone-based heuristics to help them rapidly process information on the 10-K filing date. Although we have no direct evidence that investors actually do this, we can point to anecdotal evidence that quantitative hedge funds use tone metrics calculated using machine-learning algorithms to make investment decisions.

References

Amel-Zadeh, A. and Faasse, J. (2016). The information content of 10-k narratives: comparing md&a and footnotes disclosures. *Available at SSRN 2807546*.

Andrei, D. and Hasler, M. (2015). Investor attention and stock market volatility. *The Review of Financial Studies*, 28(1):33–72.

Bali, T., Peng, L., and Shen, Y.and Tang, Y. (2014). Liquidity shocks and stock market reactions. *The Review of Financial Studies*, 27:1434–1485.

Bao, Y. and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, 60(6):1371–1391.

Barber, B. and Odean, T. (2008). All that glitters: The effect of attention on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21:785–818.

Barberis, N. and Shleifer, A. (2003). Style investing. *Journal of Financial Economics*, 68:161–199.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brown, N. C., Crowley, R. M., and Elliott, W. B. (2020). What are you saying? using topic to detect financial misreporting. *Journal of Accounting Research*, 58(1):237–291.

Bushee, B. J. and Miller, G. S. (2012). Investor relations, firm visibility, and investor following. *The Accounting Review*, 87(3):867–897.

Cohen, L. and Frazzini, A. (2008). Economic links and predictable returns. *Journal of Finance*, 63:1977–2011.

Cohen, L. and Lou, D. (2012). Complicated firms. *Journal of Financial Economics*, 104:383–400.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *Journal of Finance*, 75:1371–1415.

Crouchley, R. (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association*, 90(430):489–498.

Davis, A. K. and Tama-Sweet, I. (2012). Managers' use of language across alternative disclosure outlets: Earnings press releases versus md&a. *Contemporary Accounting Research*, 29(3):804–837.

DellaVigna, S. and Pollett, J. (2007). Demographics and industry returns. *American Economic Review*, 97:1167–1702.

Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of accounting research*, pages 123–145.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, 64(2-3):221–245.

El-Haj, M., Rayson, P., Walker, M., Young, S., and Simaki, V. (2019). In search of meaning: Lessons resources, and next steps for computational analysis of financial disclosure. *Journal of Business, Finance and Accounting*, 46(3-4):265–306.

Elliott, C. (1992). The affective reasoner: A process model of emotions in a multi-agent system. *PhD thesis, Northwestern University, The Institute for the Learning Sciences Technical Report*.

Engelberg, J. (2008). Costly information processing: Evience from earnings announcements. *Unpublished University of California San Diego working paper*.

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, 15(4):915–953.

Fiske, S. T. and Taylor, S. E. (1991). *Social cognition*. Mcgraw-Hill Book Company.

Gompers, P., Ishii, J., and Metrick, A. (2003). Corporate governance and equity prices. *The quarterly journal of economics*, 118(1):107–156.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Guo, L., Li, Frank Weikai, F., and Wei, K. J. (2020). Security analysts and capital market anomalies. *Journal of Financial Economics*, 137:204–230.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Hendershott, T., Menkveld, A. J., Praz, R., and Seasholes, M. S. (2021). Asset price dynamics with limited attention. *forthcoming The Review of Financial Studies*.

Henry, E. and Leone, A. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *The Accounting Review*, 91(1):153–178.

Hirshleifer, D., Hou, K., Teoh, S., and Zhang, Y. (2004). Do investors overvalue firms with bloated balance sheets. *Journal of accounting and economics*, 38:297–331.

Hirshleifer, D., Hsu, P., , and Li, D. (2013). Innovative efficiency and stock returns. *Journal of Financial Economics*, 107:632–654.

Hirshleifer, D., Lim, S., , and Teoh, S. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance*, 64:2289–2325.

Hirshleifer, D. and Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of accounting and economics*, 36(1-3):337–386.

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

Hong, H. and Stein, J. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *Journal of Finance*, 54:2143–2184.

Hong, H. and Stein, J. (2007). Simple models and paradigm shifts. *Journal of Finance*, 63:1207–1242.

Hong, H., Torous, W., and Valkanov, R. (2007). Do industries lead the stock market? *Journal of Financial Economics*, 83:367–396.

Hou, K. and Moskowitz, T. (2005). Market frictions, price delay, and the cross-section of expected returns. *The Review of Financial Studies*, 18:981–1020.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, 64(6):2833–2855.

Huang, X., Teoh, S. H., and Zhang, Y. (2014). Tone management. *The Accounting Review*, 89(3):1083–1113.

Huberman, G. and Regev, T. (2001). Contagious speculation and a cure for cancer: A non-event that made stock prices soar. *Journal of Finance*, 56:387–396.

Kahneman, D. (1973). *Attention and effort*, volume 1063. Citeseer.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4):237.

Lewis, C. and Young, S. (2019). Fad or future? automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, 49(5):587–615.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

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

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

Mayew, W. J., Sethuraman, M., and Venkatachalam, M. (2015). Md&a disclosure and the firm's ability to continue as a going concern. *The Accounting Review*, 90(4):1621–1651.

Mohammad, S. and Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Muslu, V., Radhakrishnan, S., Subramanyam, K., and Lim, D. (2014). Forward-looking md&a disclosures and the information environment. *Management Science*, 61(5):931–948.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Nisbett, R. E. and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall.

Peng, L. and Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80:563–602.

Read, J. (2004). Recognising affect in text using pointwise-mutual information. *Unpublished University of Sussex working paper.*

Ryans, J. (2019). Textual classification of sec comment letters. *Journal of Accounting Research, forthcoming.*

SEC (2020). How to read a 10-k.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis.* MIT press.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

University of Notre Dame (2021). Software repository for accounting and finance - `https://sraf.nd.edu/data/`.

Table 3: Summary Statistics. The table reports the mean, standard deviation (SD), and Median (Med.) for the full sample and sub-samples based on whether the firm has analysts coverage.

	Full Sample			Firms with Analyst Coverage			Firms without Analyst Coverage		
Variables	Mean	SD	Med.	Mean	SD	Med.	Mean	SD	Med.
Tone	0.79	0.71	1.00	0.81	0.69	1.00	0.70	0.80	1.00
Optimism	0.58	0.67	1.00	0.60	0.65	1.00	0.48	0.72	0.00
Specificity	1.72	0.49	2.00	1.72	0.49	2.00	1.75	0.49	2.00
Directness	1.73	0.49	2.00	1.72	0.49	2.00	1.75	0.49	2.00
Aggressiveness	0.56	0.88	0.00	0.55	0.87	0.00	0.59	0.92	0.00
AAI	3.80	0.77	4.00	3.81	0.75	4.00	3.78	0.84	4.00
Boilerplate	0.30	0.08	0.29	0.30	0.08	0.29	0.30	0.09	0.29
Number of sentences	338	226	289	354	234	300	266	170	232
Number of words	11,157	7,947	9,432	11,730	8,122	9,866	8,590	6,521	7,406
Filing date CAR (%)	0.09	5.73	-0.06	0.08	5.45	-0.05	0.14	6.82	-0.13
Book to Market	0.71	0.33	0.71	0.68	0.33	0.67	0.84	0.32	0.92
Leverage	0.18	0.21	0.11	0.19	0.21	0.13	0.13	0.20	0.05
Free Cash Flow to Assets	0.04	0.22	0.05	0.05	0.19	0.06	0.02	0.32	0.02
Acquire	0.13	0.33	0.00	0.13	0.34	0.00	0.10	0.30	0.00
Downsize	0.02	0.12	0.00	0.01	0.11	0.00	0.03	0.17	0.00
Number of Segments	1.92	4.19	1.39	2.06	4.52	1.39	1.28	2.11	1.39
Herfhindahl Index	0.01	0.05	0.00	0.01	0.05	0.00	0.00	0.05	0.00
Earnings Surprise	0.00	0.24	0.00	0.00	0.04	0.00	0.00	0.55	0.00
Earnings to Total Assets	0.01	0.29	0.03	0.01	0.27	0.03	-0.01	0.34	0.01
Earnings Volatility	0.09	0.48	0.03	0.08	0.39	0.03	0.11	0.74	0.03
Revenue	3,336	11,260	496	3,985	12,288	738	431	2,861	68
Industry Earnings Guidance	0.14	0.10	0.13	0.14	0.10	0.15	0.11	0.09	0.09

Table 4: Economic Determinants of MD&A Attributes. The table reports panel regressions of attribute classification levels for our sample of firms that have MD&A section in the 10-K reports from 2002 to 2018. One observation is one firm in one year. The dependent variables are the attribute classification levels predicted by the Naïve Bayes classifiers (Tone, Optimism, Specificity, Directness, Aggressiveness) and the Aggregate Attribute Index ($AAI_{i,t}$), which are reported in separate columns. We treat multiple the absolute value of the Aggressiveness attribute -1 so that a negative value can be interpreted as an unfavorable value. This is analogous to how Aggressiveness is used to calculate $AAI_{i,t}$. All regressions include firm and year fixed effects. Standard errors are clustered by two-digit SIC code. All regressions include firm and year fixed effects. ***, **,* respectively denote significance at the 1%, 5%, and 10% levels.

	Tone		Optimism		Specificity		Directness		$-\lvert Aggressiveness \rvert$		$AAI_{i,t}$	
Variables	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat	Coeff.	t-stat
Earnings percentile	0.0026	8.90 ***	-0.0001	-0.09	0.0001	0.56	0.0001	1.09	0.0005	3.26 ***	0.0016	6.35 ***
Free cash flow	0.0008	2.65 ***	0.0153	1.49	0.0005	3.62 ***	0.0004	3.08 ***	-0.0003	-1.00 **	0.0018	6.31 ***
Earnings volatility	-0.2649	-4.86 ***	-0.1002	-5.98 ***	-0.1036	-8.62 ***	-0.1224	-10.32 ***	0.0727	4.89 ***	-0.3245	-13.17 ***
Firm book to market	-0.0569	-2.43 **	-0.1052	-8.42 ***	-0.0151	-1.61	-0.0146	-1.58	-0.0204	-1.77 *	-0.0383	-2.00 **
Segments	0.0023	1.66 *	0.0001	0.13	0.0007	0.97	0.0006	0.95	0.0026	3.03 ***	-0.0007	-0.52
Herfhindahl index	-0.0962	-0.40	-0.0098	-0.05	-0.1206	-0.84	0.0334	0.23	0.4657	2.61 ***	-0.6232	-2.11 **
Acquire	-0.0114	-1.21	-0.0054	-0.67	0.0048	0.83	0.0142	2.48 **	-0.0019	-0.26	0.0132	1.11
Downsize	-0.0913	-3.64 ***	-0.0853	-4.82 ***	0.0017	0.14	0.0065	0.53	0.0079	0.51	-0.0693	-2.71 ***
Leverage	-0.0847	-2.67 ***	-0.2314	-10.54 ***	0.1359	8.56 ***	0.1145	7.31 ***	-0.1978	-10.07 ***	0.0815	2.50 **
Adjusted R-squared	0.051		0.022		0.029		0.022		0.051		0.042	
Observations	51,013		50,981		51,015		51,008		51,025		51,025	
Firm fixed effects	Yes		Yes		Yes		Yes		Yes		Yes	
Year fixed effects	Yes		Yes		Yes		Yes		Yes		Yes	

Table 5: Attribute Selection Random-Effects Ordered Probit Model. The table reports a coefficient estimates for a random-effects ordered probit regression of the aggregate attribute score. The dependent variable is the aggregate attribute score ($AAI_{i,t}$), which is calculated as the sum of the predicted attribute classifications from attribute-specific Naïve Bayes classifiers, i.e., $AAI_{i,t} = Tone_{i,t} + Optimism_{i,t} + Specificity_{i,t} + Directness_{i,t} - |Aggressiveness_{i,t}|$. $AAI_{i,t}$ is truncated at -4 and +4. The model includes unreported fiscal year dummies. ***, **,* respectively denote significance at the 1%, 5%, and 10% levels.

Variable	Coefficient	t-stat	
Book to Market	-0.273	-6.16	***
Earnings to Assets	-0.002	-0.20	
Leverage	0.508	6.35	***
Free Cash Flow to Assets	0.105	3.17	***
Acquire	0.162	4.37	***
Downsize	-0.006	-0.09	
Volatility	-0.522	-8.43	***
Number of Segments	-0.127	-5.53	***
Herfhindahl Index	-0.191	-0.35	
Number of Observations	33,722		
Number of Unique Firms	5,657		

Table 6: Reported Earnings and MD&A Narrative Attributes. The table reports results from a panel regression that estimates the association between reported earnings and attribute classification levels for our sample of firms that have MD&A section in the 10-K reports from 2002 to 2018. One observation is one firm in one year. The dependent variable is the net income to total asset percentile. The independent variables include the attribute classification levels predicted by the Naïve Bayes classifier (Tone, Optimism, Specificity, Directness, and |Aggressiveness|). The additional independent variables include the log of the sentences, the log of the number of words, log of the firm book to market ratio, earnings five-year earnings volatility, the log of revenues, and the percentage of firms in the same two-digit SIC code that provide earnings guidance. The specification uses a Heckman correction to control for the endogenous selection of narrative content. Standard errors are clustered by two-digit SIC code. All regressions include firm and year fixed effects. ***, **,* respectively denote significance at the 1%, 5%, and 10% levels.

	Analyst Coverage			No Analyst Coverage		
Variables	Coeff.	t-stat		Coeff.	t-stat	
Tone	1.726	5.07	***	1.259	3.77	***
Optimism	1.556	4.32	***	1.353	3.84	***
Specificity	-0.343	-0.73		-0.064	-0.09	
Directness	1.045	2.10	**	-0.071	-0.11	
|Aggressiveness|	-0.880	-2.78	***	-0.316	-0.73	
Boilerplate	0.505	0.17		3.607	1.12	
Inverse Mills ratio	-0.547	-2.90	***	-0.081	-0.40	
Log(Number of sentences)	7.211	1.98	**	11.341	2.90	***
Log(Number of words)	-10.420	-2.84	***	-14.600	-3.76	***
Log(Firm book to narket)	-20.192	-8.68	***	-15.070	-10.47	***
Earnings volatility	-4.053	-2.32	**	-1.854	-1.08	
Log(Revenues)	7.725	6.99	***	6.911	7.17	***
Industry Earnings Guidance	9.013	1.50		-4.354	-0.47	
Adjusted R-squared	0.279			0.202		
Observations	30,016			20,234		
Firm fixed effects	Yes			Yes		
Year fixed effects	Yes			Yes		

Table 7: Filing Date Cumulative Abnormal Returns and MD&A Attributes. The table reports regressions of three-day cumulative abnormal returns (-1, +1) on filing dates that follow a previous earnings announcement date for our sample of firms that have MD&A section in 10-K reports from 2002 to 2018. The dependent variable is the three-day cumulative abnormal return on the filing date. The primary independent variables reflect textual attributes for tone, optimism, specificity, vagueness, aggressiveness, and the proportion of boilerplate sentences. The specification uses a Heckman correction to control for the endogenous selection of narrative content. All regressions include firm and year fixed effects. Standard errors are clustered by firm and year. Robust standard errors are calculated using Huber-White Sandwich estimators and double clustered by firm and year. We suppress reporting of the additional explanatory variables that are not associated with textual attributes. ***, **,* respectively denote significance at the 1%, 5%, and 10% levels.

	Firms with Analyst Coverage			Firms without Analyst Coverage		
	(1)			(1)		
Variables	Coeff.	t-stat		Coeff.	t-stat	
Tone	0.1381	2.00	**	0.31938	2.13	**
Optimism	-0.0825	-1.11		0.33554	2.07	**
Specificity	-0.0286	-0.27		0.03891	0.15	
Directness	0.0137	0.12		0.03039	0.11	
Passive Aggressive (PA)	-0.0713	-1.62		0.17010	1.22	
Boilerplate	-0.8276	-1.56		5.72620	4.15	***
Inverse Mills Ratio	0.0040	0.12		0.24656	2.67	***
Log(Number of Sentences)	0.8276	1.57		-0.30346	-0.28	
Log(Number of Words)	-0.8414	-1.62		0.47939	0.46	
Log(Firm Book to Market)	0.1606	2.19	**	0.12966	0.59	
Earnings to Total assets	-1.2909	-3.37	***	0.41313	1.14	
Earnings Volatility	0.1667	2.54	**	0.14179	2.63	***
Log(Revenues)	0.0313	1.48		0.01555	0.21	
Adjusted R squared	0.0069			0.0069		
Observations	22,621			5,002		

Table 8: 60-Day Buy-And-Hold Abnormal Returns and MD&A Attributes for Consistent and Inconsistent Tone Messaging. The table reports regressions of 60-day buy-and-hold returns (BHARs) (+2, +60) relative to filing dates that follow a previous earnings announcement date for our sample of firms that have MD&A section in 10-K reports from 2002 to 2018. Panel A reports the results for firms that have released MD&A sections with tone messages that are consistent with unexpected earnings, e.g., positive tone is used to describe unexpectedly good earnings news. Panel B reports the results for firms that have released MD&A sections with tone messages that are inconsistent with unexpected earnings, e.g., positive tone is used to describe unexpectedly bad earnings news. The dependent variable is the three-day cumulative abnormal return on the filing date. The primary independent variables reflect textual attributes for tone, optimism, specificity, vagueness, aggressiveness, and the proportion of boilerplate sentences. The specification uses a Heckman correction to control for the endogenous selection of narrative content. All regressions include year fixed effects. Robust standard errors are calculated using Huber-White Sandwich estimators and double clustered by firm and year. We suppress reporting of the additional explanatory variables that are not associated with textual attributes. ***, **,* respectively denote significance at the 1%, 5%, and 10% levels.

	Firms with Analyst Coverage			Firms without Analyst Coverage		
Variables	Coeff.	t-stat		Coeff.	t-stat	
Panel A: Consistent Tone Messaging Relative to Earnings Surprise						
Optimism	0.9450	2.57	**	0.9180	1.49	
Specificity	0.4737	0.83		1.2858	1.03	
Directness	-0.4154	-0.74		0.9904	0.68	
Passive	-0.9012	-2.03	*	0.2352	0.30	
Aggressive	-1.2711	-0.88		2.7647	0.67	
Boilerplate	-6.5711	-2.18	**	5.6223	0.92	
Inverse Mills Ratio	0.4634	2.87	**	0.1861	0.65	
Log(Number of Sentences)	-0.8010	-0.27		0.8256	0.16	
Log(Number of Words)	0.0529	0.02		-0.0442	-0.01	
Adjusted R squared	0.0241			0.0417		
Panel B: Inconsistent Tone Messaging Relative to Earnings Surprise						
Optimism	1.2722	3.14	***	0.8192	1.24	
Specificity	-0.0925	-0.13		-1.1827	-0.58	
Directness	-0.7012	-1.04		0.6256	0.35	
Passive	0.6103	1.80	*	-0.1707	-0.16	
Aggressive	1.2717	0.45		5.2253	1.26	
Boilerplate	-3.7372	-1.09		5.7137	0.93	
Inverse Mills Ratio	-0.4191	-1.90	*	-0.8400	-2.05	*
Log(Number of Sentences)	-5.2872	-1.93	*	-3.3335	-1.01	
Log(Number of Words)	4.4837	1.81	*	2.9820	0.92	
Adjusted R squared	0.0366			0.0526		

A Appendix - Naïve Bayesian Classification

A.1 Building the Naïve Bayes Classifiers

We collect our data using REDCap which is a secure web application for building and managing on line surveys and databases.[22] A key feature of REDCap is that respondents are able to access a survey from virtually any device with internet capabilities. This allow us to solicit a large number of potential respondents that included members of our LinkedIn networks and students enrolled in graduate accounting and finance courses. Based on private correspondences, we were able to survey individuals both in academia and with a wide variety of professional backgrounds and expertise in financial services, auditing, consulting, asset management and financial regulation. Although it would be interesting to analyze classification differences across respondents, we did not collect demographic data because we believe that anonymity maximizes response rates. The final survey reflects 283 respondents that scored and average of 44.5 sentences each.[23]

Crowd sourcing the data collection phase of our project has several advantages. First, it ensures that the results are not subject to the personal biases of a small set of individuals. Second, since respondents can score as many sentences as they want and there are no limits on how often a respondent can access the survey tool, the risk of "classification fatigue" is strongly limited. These features increase our ability to collect a large amount of data in a timely and accurate manner.

An example of the REDCap survey tool is displayed in Figure 1. Each time a respondent classifies a sentence, the survey tool accesses a file that contains a set of preselected sentences and randomly displays one of them for the respondent. Below each sentence, the survey asks respondents to answer the following questions:

- *Tone.* Is the tone of the sentence positive or negative?
- *Optimism.* Does the sentence reflect either an optimistic or pessimistic view by the company?
- *Specificity.* Does the sentence provide specific details that are supported by objective facts?
- *Evasiveness.* Does the sentence appear to reflect an attempt by the company to be non-responsive?

[22] It has reporting functions that provide summary statistics and includes automated export procedures for downloading the data into Excel and common statistical packages like SAS, Stata, and R.

[23] The survey is designed to remain open indefinitely and the number of classified sentences grows over time.

- *Aggressiveness.* Does the sentence reflect a strongly worded assertion?
- *Boilerplate.* Does the sentence reflect a scripted response?

Respondents either select a classification level, as shown in Table 1, or can report the attribute "Not Applicable" if none of the linguistic attribute seems plausible or applicable. The final question asks the respondent to indicate whether the sentence should be classified as boilerplate. Throughout the paper, we refer to "Evasiveness" as "Directness" to facilitate its interpretation in the context of other attributes like Optimism and Specificity both of which tend to have favorable connotations.

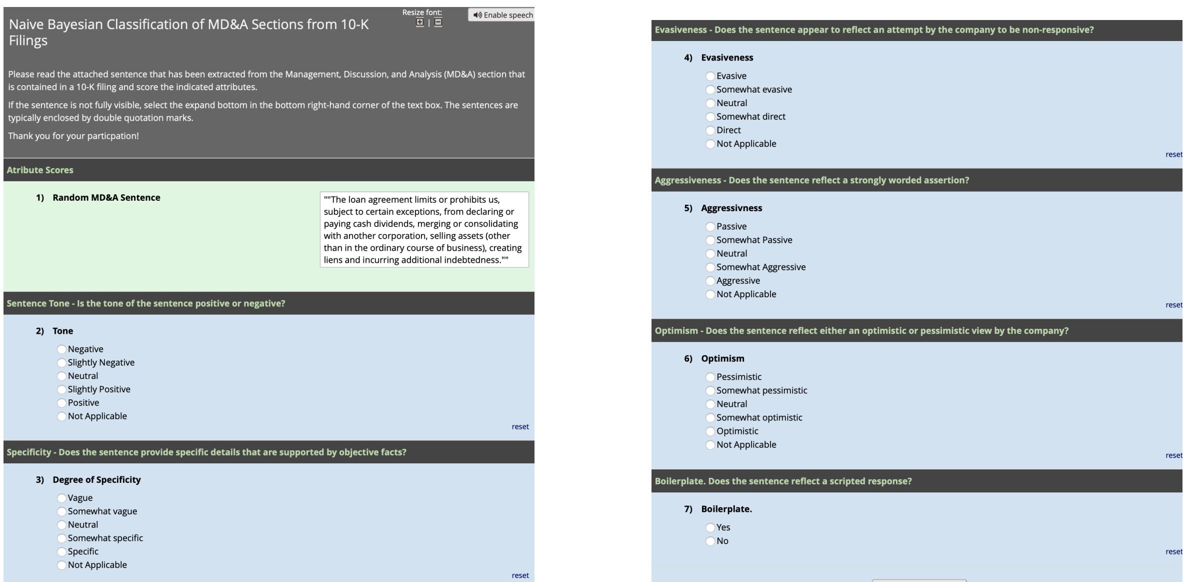

Figure 1: Example sentence from REDCap survey.

A.2 Characteristics of the Training Sample

The sentences that are *scored* with the REDCap survey tool are used to train Naïve Bayes classifiers (NBCs) for each attribute. Panel A of Table A.1 reports the classification frequency distributions for each attribute. The attribute frequency distributions for the training sample are depicted graphically in Figure 2. The Tone distribution is reasonably symmetric with 7.04% negative and 8.86% positive and 14.44% somewhat negative and 13.37% somewhat positive. Aggregating positive (+1 and +2) and negative (-2 and -1) Tone classifications, we see that 21.58% are negative and 27.81% are positive. Our sample frequencies of negative and positive tone are higher than those reported by Li [2010] who reports negative and positive frequency rates of 17.82% and 19.59%, respectively.

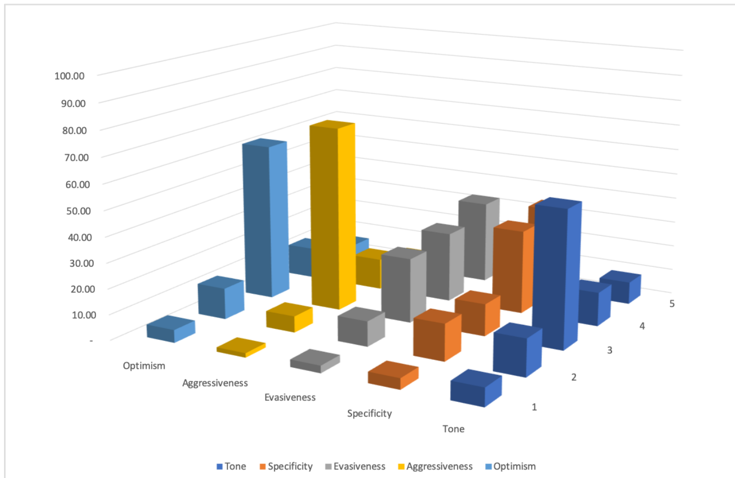

Figure 2: Sentence-Level Attribute Frequency Distributions for Training Sample

The distribution for Optimism also is symmetric. The main difference is that there are more neutral responses for Optimism (62.28%) relative to Tone (53.66%). This finding seems sensible because not all positive statements are optimistic, but most optimistic statements would also be considered positive. The frequency distributions for Specificity and Directness are skewed toward more specific and direct classifications. For example, aggregating the classifications into three levels $\{(-2, -1); (0); (+1, +2)\}$, we find that 67.70% (33.13% + 34.56%) and 60.60% (27.72% + 32.87%) of all sentences in the training sample are classified as Direct and Specific. We conjecture that this is attributable to the objective nature of the MD&A sections which tend to be devoid of emotional nuance. For the Aggressiveness attribute, there is a slight tendency for sentences to be classified as aggressive rather than passive, but the vast majority (73.03%) are classified as having neutral content.

Panel B of Table A.1 reports that within-sample "balanced" accuracy rates range from 68.18% to 88.04% with most attribute classification categories exceeding 75% accuracy rates. Balanced accuracy is the average of sensitivity (i.e., the true positive classification rate) and specificity (i.e., 1 - false positive classification rate).[24] This indicates that the model does an excellent classification job within-sample.

Panel C of Table A.1 reports that the mean balanced accuracy from 1,000 repeated 10-fold cross validation tests. As one would expect from any out-of-sample test, the balanced accuracy decreases relative to the within-sample results in Panel A. We find that 16 of 25 classifiers have mean balanced accuracy levels that exceed 55%. Only one out-of-sample classifier has

[24]The true positive classification rate is the ratio of true positive responses to the sum of true positives and false negative responses. The false positive classification rate is the ratio of true negative responses to the sum of false positive responses and true negative responses.

a balanced accuracy rate below 50%. The aggressive classification level (-2) for the Aggressiveness attribute is 49.48%, which is effectively a coin toss. This is likely attributable to the fact that only 1.70% of the sentences are classified as Aggressive. The relative small number of sentences likely renders the classifier somewhat noisy. We observe similar behavior for the evasive classification level (-2) of the Directness attribute. In this case, the probability of correctly classifying an evasive sentence only is 50.48%, which once again, is another coin toss. From a practical standpoint, the relatively low frequency rates are expected as one would not anticipate many firms providing aggressive or evasive content in an MD&A.

To address the possible cause for the decrease in out-of-sample balanced accuracy, we consider whether the lower accuracy rates could be attributable to relatively minor misclassifications between (-2, -1) and (+1, +2) classification levels. To do this, we calculate an Aggregate Classification Rate (ACR) that is based on a broader notion of correct classification that is analogous to specificity. We treat classification levels (-2, -1), (0), and (+1,+2) as three separate categories.[25] The ACR is calculated as:

$$ACR = \frac{CorrClass}{CorrClass + InCorrClass}, \tag{11}$$

where,[26]

$$\begin{aligned} CorrClass &= TrueNeg + TrueNeut + TruePos \\ InCorrClass &= FalseNeg + FalseNeut + FalsePos. \end{aligned}$$

For example, this approach treats a "Positive" (+2) sentence that was classified by the NBC as "Somewhat Positive" (+1) as a correct classification. In other words, ACR does not differentiate between unambiguous classification levels (-2 and +2) and "Somewhat" classification levels (-1 and +1).

Panel A reports that the within-sample ACRs range from 79.01% for *Tone* to 84.37% for *Aggressiveness*. This suggests that the lower within-sample balanced accuracy rates are largely attributable to relatively minor misclassifications between (-2, -1) and (+1, +2) classification levels. The same conclusion holds for the out-of-sample cross validations tests in Panel B.

[25] This approach is in the spirit of Li [2010] who aggregates negative and neutral responses into a single category.

[26] $TrueNeg$ is the number of sentences that have actual and predicted classification levels of -2 and -1, $TrueNeut$ is the number of sentences that have actual and predicted classification levels equal to 0, and $TruePos$ is the number of sentences that have actual and predicted classification levels of +1 and + 2. $FalseNeg$ is the number of sentences that have positive classifications but are predicted to have negative classifications, $FalseNeut$ is the number of sentences that have non-neutral classifications but are predicted to have neutral classifications, and $FalsePos$ is the number of sentences that have negative classifications but are predicted to have positive classifications.

Although the ACRs decrease relative to the within-sample ACRs, accuracy rates remain relatively high, ranging from 56.30% for Directness to 84.53% for Aggressiveness. Based on these results, we use aggregated classifications in our empirical analyses.

A.3 Predicting Textual Attributes

Using the NBCs trained above, we then classify 29,677,476 sentences identified in 118,899 MD&As along the five narrative attributes (Tone, Optimism, Specificity, Directness, and Aggressiveness) and predict whether each sentence should be considered boilerplate. Our final sample covers the period from 2001 through 2018. We identify 8,912,118 sentences classified as boilerplate which are then removed, resulting in a final sample of 20,765,358 sentences.

Table A.2 provides the predicted classification frequency distributions for each attribute. Panel A reports sentence-level frequency distributions by attribute which are also depicted graphically in Figure 3. Casual observation indicates that the predictions shown in Figure 3 have similar qualitative properties of the training sample shown in Figure 2. There is, however, a tendency for all five predicted attributes to display more right skewness. The increase in skewness is most pronounced for Specificity and Directness. By contrast, the distributions of Tone, Aggressiveness, and Optimism remain relatively symmetric. After comparing the frequency distributions for the training and final samples, we conclude that the training sample is representative of the full corpus of MD&A sentences.

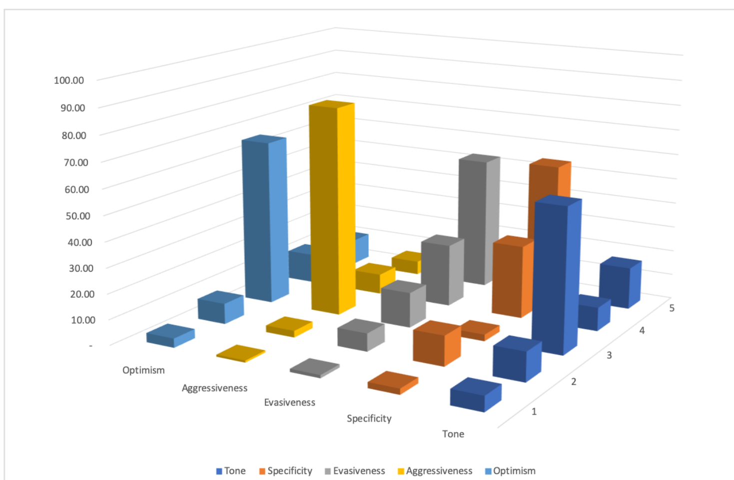

Figure 3: Sentence-Level Attribute Frequency Distributions for the final sample.

Since documents are the unit of analysis,[27] we estimate NBC classification scores on document-level word count vectors. Working with word counts at the document-level helps un-

[27]All the empirical tests discussed in the paper have firm-year observations as the statistical unit of analysis.

cover possible textual nuance because they are far less sparse then sentence-level word counts. An alternative approach is to classify individual sentences and aggregate the information to create document-level metrics. For instance, one can compute the analog to the document-level attribute scores as the mean attribute score for all non-boilerplate sentences in a given document. Again, we do not pursue this approach because sentence-level word counts are sparsely populated and a large number of sentences are classified as Neutral.

Panel B of Table A.2 reports document-level frequency distributions by attribute which are also depicted graphically in Figure 4. The number of documents that are classified as being at least Somewhat Positive is 60.19%, while only 11.81% are at least Somewhat Negative. Given the relatively symmetric rate of positive and negative sentences, one may find this result somewhat surprising. One possible explanation is that, on average, managers may attempt to "spin" MD&A discussions to explain performance in a favorable manner and that such nuance is more readily detected in the aggregate. This is reinforced by the relative high number of documents that are classified as optimistic (47.85%).

We find that 95.84% and 95.22% of all MD&A sections are classified as being at least Somewhat Specific and Somewhat Direct, respectively. This compares to analogous rates at the sentence-level of 83.36% and 77.74%. High classification scores for the Specific and Direct attributes is hardly surprisingly given the factual content of MD&A filings. Finally, we note that 33.17% of all MD&A sections are classified as Aggressive, which is six-times higher than what was observed at the sentence-level (5.52%).

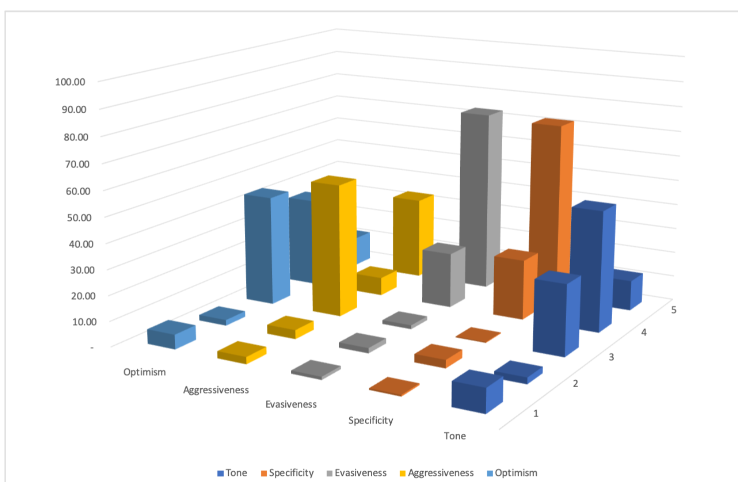

Figure 4: Document-Level Attribute Frequency Distributions for the final sample.

Table A.1: Attribute Classification Frequency Distributions and Naïve Bayes Classifier Accuracy Rates. Panel A reports the frequency distribution of attribute classifications based on 8,140 non-boilerplate survey responses. Not applicable classifications (+3) are not reported. The first five columns report the actual classification levels (-2, -2, 0, +1, +2). Panel B reports balanced accuracy and aggregate classification rates (ACRs) for the full survey sample. The first five columns report balanced accuracy rates for different classification levels. The sixth column is the ACR. ACRs is an accuracy metric that collapses negative (-2 and -1), neutral (0), and positive (+1 and +2) classification levels into three groups. Panel C reports the mean balanced accuracy and ACRs for a sample of 1,000 10-fold cross validation tests. All p-values that test the hypothesis that the mean balanced accuracy rates and ACRs are statistically distinct from 0.5 are rejected at p-values that are close to zero.

	Classification Level					
Attribute	-2	-1	0	+1	+2	ACR
Panel A. Attribute Classification Frequency Distribution (%)						
Tone	7.04	14.44	53.66	13.37	8.86	
Specificity	4.23	14.39	12.63	33.13	34.56	
Directness	2.85	9.84	25.50	27.72	32.87	
Aggressiveness	1.70	6.42	73.03	12.38	4.64	
Optimism	5.01	12.32	62.28	12.37	6.18	
Panel B. Within Sample Aggregate and Balanced Accuracy Rates (%)						
Tone	88.04	77.12	76.50	79.48	80.31	79.01
Specificity	86.06	75.31	77.56	73.15	77.09	84.37
Directness	84.26	78.17	72.65	74.05	74.84	80.86
Aggressiveness	68.18	81.86	73.54	77.49	77.84	81.92
Optimism	87.14	78.20	76.70	79.35	80.22	80.73
Boilerplate						66.83
Panel C. 10-Fold Cross Validation Aggregate and Balanced Accuracy Rates						
Tone	64.30	60.19	64.84	56.18	61.24	79.25
Specificity	56.10	57.24	56.92	54.82	65.22	59.23
Directness	50.83	55.88	54.77	54.28	61.65	56.30
Aggressiveness	49.48	51.97	54.99	54.89	52.58	84.53
Optimism	62.37	60.78	63.21	58.58	57.81	83.46
Boilerplate						62.80

Table A.2: Attribute Classification Frequency Distributions for Full Sample. This table reports the frequency distribution of attribute classifications for the full sample of 20,765,358 sentences from MD&A section of 10-K filings from 2001 to 2018.

	Classification Level Frequency (%)				
Attribute	-2	-1	0	+1	+2
Panel A. Sentence-Level Attribute Classification Frequency Distribution					
Tone	5.89	11.45	56.48	9.38	16.80
Specificity	2.26	11.78	2.59	28.99	54.37
Directness	1.22	7.11	13.93	24.92	52.82
Aggressiveness	0.78	2.64	82.98	8.08	5.52
Optimism	3.63	8.20	65.86	11.86	10.45
Panel B. Document-Level Attribute Classification Frequency Distribution					
Tone	9.29	2.52	27.99	48.05	12.14
Specificity	0.61	3.22	0.34	24.04	71.80
Directness	1.19	2.05	1.54	22.13	73.09
Aggressiveness	2.79	3.58	53.14	7.31	33.17
Optimism	5.76	2.26	44.12	35.84	12.01

B Appendix - Variable Definitions and Description

Dependent Variables

Variable	Description
Tone	Naïve Bayes classification score for Tone based on the document-level classification of the MD&A for firm j in year t.
Optimism	Naïve Bayes classification score for Optimism based on the document-level classification of the MD&A for firm j in year t.
Specificity	Naïve Bayes classification score for Specificity based on the document-level classification of the MD&A for firm j in year t.
Directness	Naïve Bayes classification score for Directness based on the document-level classification of the MD&A for firm j in year t.
Aggressiveness	Naïve Bayes classification score for Aggressiveness based on the document-level classification of the MD&A for firm j in year t.
|Aggressiveness|	The absolute value of Aggressiveness score.
$AAI_{i,j}$	Aggregate Attribute Index as Tone + Optimism + Specificity + Directness-|Aggressiveness|, truncated at +4 and -4.
Earnings percentile	Ratio of net income to lagged total assets percentile.
Analyst coverage	Dummy variable that takes value 1 if a firm has analyst coverage in the I/B/E/S summary history database and zero otherwise.
CAR	Three-day cumulative abnormal return based on market model.
Unexpected earnings	Actual earnings minus expected earnings scaled by market capitalization of equity.
Expected earnings	Expected earnings are calculated as the I/B/E/S consensus forecast when available or a fourth-quarter random walk model when an analyst consensus is unavailable.

Independent Variables

Variable	Description
Earnings percentile	Return on assets (ROA) percentile. Percentiles are based on the full sample.
Firm book to market	Firm book to market is calculated as the ration of total assets to enterprise value.
Enterprise value	Enterprise Value is calculated as the market capitalization of equity (shares outstanding × price per share) plus total liabilities.
Leverage	Leverage is calculated as the ratio of long-term debt to total assets.
Free cash flow	Free cash flow is calculated as the ratio of free cash flow to lagged total assets where free cash flow is defined as earnings before interest, taxes, depreciation and amortization (EBITDA) minus capital expenditures.
Acquire	Acquire is a dummy variable for an increase in firm size based on total assets. It takes value of 1 the ratio of total assets to lagged total assets is greater than 1.33 and 0 otherwise.
Downsize	Downsize is a dummy variable for a decrease in firm size based on total assets. It takes value of 1 if the ratio of total assets to lagged total assets is less than 0.67 and 0 otherwise.
Earnings volatility	Earnings volatility is calculated as the standard deviation of return on assets for the last five years.

Segments	Sum of Compustat segment identifiers (SID) as $\log(1 + \text{Number of segments})$.
Herfhindahl index	Herfindahl index for firm j at year t is calculated as the ratio of revenues for firm j relative to industry average revenues based on the firm j's two-digit SIC.
Return on assets	Return on assets is calculated as net income scaled by total assets.
Boilerplate	Boilerplate is calculated as the percentage of sentences in the MD&A of firm j in year t that were classified as Boilerplate by the related Naïve Bayes classifier.
Log(Number of sentences)	The natural logarithm of the number of sentences in the MD&A of firm j at year t.
Log(Number of words)	The natural logarithm of the number of words in the MD&A of firm j at year t.
Industry Earnings Guidance	Percentage of firms in the same two-digit SIC providing earnings guidance over the past four quarters.