NBER WORKING PAPER SERIES

HOW TO TALK WHEN A MACHINE IS LISTENING: CORPORATE DISCLOSURE IN THE AGE OF AI

Sean Cao
Wei Jiang
Baozhong Yang
Alan L. Zhang

Working Paper 27950
http://www.nber.org/papers/w27950

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

The authors have benefitted from discussions with Emilio Calvano (discussant), Kathleen Hanley (discussant), Tim Loughran, and Song Ma, and comments and suggestions from participants in seminars and conferences at Georgia State, Peking University, Utah, and the NBER Economics of Artificial Intelligence Conference. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI
Sean Cao, Wei Jiang, Baozhong Yang, and Alan L. Zhang
NBER Working Paper No. 27950
October 2020
JEL No. G14,G30

ABSTRACT

This paper analyzes how corporate disclosure has been reshaped by machine processors, employed by algorithmic traders, robot investment advisors, and quantitative analysts. Our findings indicate that increasing machine and AI readership, proxied by machine downloads, motivates firms to prepare filings that are more friendly to machine parsing and processing. Moreover, firms with high expected machine downloads manage textual sentiment and audio emotion in ways catered to machine and AI readers, such as by differentially avoiding words that are perceived as negative by computational algorithms as compared to those by human readers, and by exhibiting speech emotion favored by machine learning software processors. The publication of Loughran and McDonald (2011) is instrumental in attributing the change in the measured sentiment to machine and AI readership. While existing research has explored how investors and researchers apply machine learning and computational tools to quantify qualitative information from disclosure and news, this study is the first to identify and analyze the feedback effect on corporate disclosure decisions, i.e., how companies adjust the way they talk knowing that machines are listening.

Sean Cao
Georgia State University
scao@gsu.edu

Wei Jiang
Graduate School of Business
Columbia University
3022 Broadway, Uris Hall 803
New York, NY 10027
and NBER
wj2006@columbia.edu

Baozhong Yang
J. Mack Robinson College of Business
35 Broad Street, Suite 1243,
Atlanta GA 30303
bzyang@gsu.edu

Alan L. Zhang
J.Mack Robinson College of Business
35 Broad Street NW, Suite 1242
Atlanta, GA 30303
lzhang27@gsu.edu

I. Introduction

The annual report (and other regulatory filings) is more than a legal requirement for public companies; it provides an opportunity to communicate financial health, to promote the culture and brand, and to engage with a full spectrum of stakeholders. How those readers process the wealth of information affects their perception of, and hence participation in, the business in significant ways. Warren Buffet's annual letters to shareholders in Berkshire Hathaway's annual reports showcase Corporate American writing at its best. "Be fearful when others are greedy and greedy when others are fearful," Buffett wrote in the 2007 report. "When it's raining gold, reach for a bucket, not a thimble." He added in 2009. That is an entire business philosophy in 20 words.

However, there are many reasons why the Buffett writing is an envy but is hard to emulate. Added to such a list of reasons is the evolving potential readership in the age of AI (Artificial Intelligence). More and more companies realize that the target audience of their mandatory and voluntary disclosures no longer consists of just human analysts and investors. A substantial amount of buying and selling of shares are triggered by recommendations made by robots and algorithms which process information with machine learning tools and natural language processing kits.[1] Both the technological progress and the sheer volume of disclosure make the trend inevitable.[2] Companies who wish to accomplish the desired outcome of communication and engagement with stakeholders need to adjust how they talk about their finances, brands, and make

[1] For example, Kensho (acquired by S&P in 2018 in the largest AI-driven acquisition deal at the time) developed an algorithm named Warren (after Warren Buffett) that provides a simple interface allowing investors to ask complex questions in plain English and provide answers by searching through millions of market data points. (Source: "Wall Street Tech Spree: With Kensho Acquisition S&P Global Makes Largest A.I. Deal in History," Antoine Gara, Forbes, March 6, 2018). A leading hedge fund, the Man Group, has begun to manage substantial portions of its assets using AI and algorithmic trading. (Source: "The Massive Hedge Fund Betting on AI," Adam Satariano and Nishant Kumar, Bloomberg, September 27, 2017.)

[2] Cohen, Malloy, and Nguyen (2020) document that the length of 10-K increases by five times from 2005 to 2017, and the number of textual changes over previous filings increases by over 12 times.

forecasts in the age of AI. In other words, they should heed to the unique logic and techniques underlying the rapidly evolving language- and sentiment-analysis facilitated by large-scale machine-learning computation, for example, automated computational processes that identify positive, negative and neutral opinions in a whole corpus of a firm disclosure that is beyond processing ability of human brains. While the literature is catching up with and guiding investors' rising aptitude to apply machine learning and computational tools to extract qualitative information from disclosure and news, there has not been an analysis exploring the *feedback effect*, i.e., how companies adjust the way they talk knowing that machines are listening. This paper fills this void.

Our analysis starts with a diagnostic test that connects the expected extent of AI readership for a company's SEC filings on EDGAR (measured by *Machine Downloads*), and how machine-friendly the company composes its disclosure (measured by *Machine Readability*). The first variable *Machine Downloads* is constructed, using historical information, by tracking IP addresses that conduct downloads in batches. We deem *Machine Downloads* a proxy for AI readership, both because machine request is a precursor and a necessary condition for machine reading, and because the sheer volume of machine downloads makes it unlikely for them to be processed by human readers alone. The second variable builds on the five elements, identified by the recent and burgeoning literature (see Section 2), as affecting the ease for machine parsing, scripting, and synthesizing.

We show that, in the cross-section of filings with firm and year fixed effects, a one standard deviation change in expected machine downloads is associated with 0.24 standard deviation increase in the *Machine Readability* of the filing. On the other hand, other (non-machine) downloads do not bear any meaningful correlation with machine readability validating *Machine Downloads* as a proxy for machine readership. We further validate that *Machine Downloads* and

Machine Readability are reasonable proxies (for the presence of machine readership and the ease for machines to process) by showing that trades are quicker to follow after a filing becomes public when *Machine Downloads* is higher, with even stronger interactive effect with *Machine Readability*. Such a result also demonstrates the real impact of machine-process on information dissemination.

After establishing a positive association between a high AI reader base and more machine-friendly disclosure documents, we further explore how firms manage "sentiment" and "tones" perceived by machines. It is well-documented that corporate disclosures attempt to strike the right tones with (human) readers by conveying positively-biased sentiments and favorable tones without being explicitly dishonest or overtly noncompliant (Loughran and McDonald 2011, Kothari, Shu, and Wysocki 2009). Hence, we expect a similar strategy catered to machine readers. While researchers and practitioners had long relied on the Harvard Psychosociological Dictionary (especially the Harvard-IV-4 TabNeg file) to construct "sentiment" as perceived by (mostly human) readers by counting and contrasting "positive" and "negative" words, the publication of Loughran and McDonald (2011, "LM" hereafter) presents an instrumental event to test our hypothesis pertaining to machine readers. This is because not only the paper presented a specialized finance dictionary of positive/negative words and words that are informative about liability and uncertainty, but also the word lists that came with the paper has served as a leading lexicon for algorithms to sort out sentiments in both the industry and academia.[3] The differences in both the timeline and the context of the new dictionary allow us to identify the impact of AI readership on sentiment management by corporations.

[3] The LM dictionaries have had a far-reaching influence in the academic literature, e.g., see our discussion of the literature using the LM dictionary at the end of the introduction. For examples of industry uses, see "Natural Language Processing in Finance: Shakespeare Without the Monkeys," July 2019, Man Group, and "NLP in the Stock Market," Roshan Adusumilli, February 2020, Medium.com.

As a first step, we establish that firms which expect high machine downloads avoid LM-negative words but only post 2011 (the year of publication of the LM dictionary). Such a structural change is absent with respect to words deemed negative by the Harvard Dictionary (which has served human readers for a long time). As a result, the difference, *LM – Harvard Sentiment*, follows the same path as the *LM Sentiment*. For a tighter identification, we further confirm a parallel pre-trend in the *LM – Harvard Sentiment* between firms with high and low (top and bottom terciles of) machine downloads up to 2010. Post-2011 saw a clear divergence where the "high" group significantly reduce their uses of negative words from the LM Dictionary as opposed to those from the Harvard Dictionary, relative to the "low" group. Given the quasi-randomness of the exact timing of publication, the change in the sentiment expression is more likely to be attributed to firms' catering to its AI readers than an alternative hypothesis that the publication was a side show of a pre-existing and continuing trend.

Loughran and McDonald (2011) developed multiple additional dictionaries of "tone" words aiming at capturing a richer set of annotations of a financial document, including dictionaries of litigious, uncertain, weak modal, and strong modal words. The authors show that the prevalence of words in each category predict firm outcomes such as legal liability and reaction from the capital markets. We find that firms with higher expected machine readership became more averse to words from these dictionaries following the Loughran and McDonald (2011) publication. The combined results suggest that managers revise their corporate disclosure in consideration of multi-dimensional effects of their words to the eyes of the machines.

While our analyses thus far focus on the textual information, the application of the underlying theme (i.e., "how to talk when a machine is listening") to the speech setting serves as an out-of-sample test beyond the textual setting. Earlier work by Mayew and Ventakachalam

(2012) find that managers' vocal expressions can convey incremental information valuable to analysts covering the firm. Given that machine learning software makes vocal analytics more and more effective, managers should also recognize the possibility that their speeches need to impress machines as well as humans. Applying a popular pre-trained machine learning software to extract two emotional features well-established in the psychology literature, valence and arousal (correspond to positivity and excitedness of voices) on managerial speeches in conference calls, we find that managers of firms with higher expected machine readership exhibit more positivity and excitement in their vocal tones, justifying the anecdotal evidence that managers increasingly seek professional coaching to improve their vocal performances along the quantifiable metrics.[4]

Our study builds on an expanding literature on information acquisition and dissemination via SEC filings downloads,[5] opting in a new angle on the consequences of machine downloads and potentially machine processing. Our paper also contributes to the rapidly growing literature on textual analysis with a central theme that qualitative information from, and writing quality of, texts predicts asset returns and corporate performance.[6] The computational textual analyses have been steadily advanced by more modern machine learning techniques,[7] and have been extended to

[4] Sources: "Listening Without Prejudice: How the Experts Analyze Earnings Calls for Lies, Bluffs, and Other Flags", Sterling Wong, Minyanville, April 18, 2012. "How to listen for the hidden data in earnings calls", Alina Dizik, Chicago Booth Review, May 25, 2017.

[5] Recent studies analyzing downloads of SEC filings include Bernard, Blackburne, and Thornock (2020), Cao, Du, Yang, and Zhang (2020), Chen, Cohen, Gurun, Lou, and Malloy (2020), and Crane, Crotty, and Umar (2020).

[6] Tetlock (2007), Tetlock, Saar-Tsechanksy, and Macskassy (2008), and Hanley and Hoberg (2010) pioneered applying psychological dictionaries to financial texts to given content to sentiments. LM (2011) developed capital-market specific dictionaries which have since been applied to large-scale computation of tones and sentiment in financial texts, e.g., Dow Jones newswires (Da, Engelberg, and Gao, 2011), New York Times financial articles (Garcia, 2013), 10-K and IPO prospectuses (Jegadeesh and Wu, 2013), corporate press releases (Ahern and Sosyura, 2014), earnings conference calls (Jiang, Lee, Martin, and Zhou, 2019), and all wired news from Factiva (Huang, Tan, and Wermers, 2020). Hwang and Kim (2017) directly connect the writing quality of filings to valuation in the context of close-end funds. See also the survey article Loughran and McDonald (2016).

[7] Applications of more recent techniques in finance research include support vector regressions (Manela and Moreira, 2017), word embedding and Latent Dirichlet Analysis (Li, Mai, Shen, and Yan, 2020; Hanley and Hoberg, 2019; Cong, Liang, and Zhang, 2019), and neural networks (Chen, Wu, Yang, 2019). See also the survey article Cong, Liang, Yang and Zhang (2020).

non-text data such as the audios of conference calls (Mayew and Ventakachalam, 2012) and videos of startup pitch presentations (Hu and Ma, 2020). Our study departs from the existent literature as we explore managerial disclosure strategies in response to the growing presence of AI analytical tools in both the industry and academia.

Our study thus connects to a distinct literature on the "feedback effect," that is, while the financial markets reflect firm fundamentals, the market perception also influences manager's information set and decision making (see a survey by Bond, Edmans, and Goldstein, 2012). Our study uncovers a novel "feedback effect" of machine learning about firm fundamentals on corporate decisions in the era of AI. As long as the encoded rules are not completely opaque—because such rules are transparent, observable, or reverse-engineerable to at least some degree, agents who are impacted by the decisions have the incentive to manipulate the inputs to machine learning in order to game at a more desirable outcome. Though a relation between metrics and behavior is not new,[8] it is fairly recent that the machine learning community formalizes the matter as one of "strategic classification" (Hardt, Megiddo, Papadimittriou, and Wootters, 2016; Dong, Roth, Schutzman, and Waggoner, 2018; Milli, Miller, Dragan, and Hardt, 2019). We present the first empirical evidence of the feedback effect from algorithmic assessment to corporate behavior.[9] While some adaptive behavior, such as making disclosure more machine-reading friendly, is innocuous or even welcome, other algorithm-induced changes, such as the expression of sentiment, highlight the increasing challenge on machine learning to be "manipulation proof" in that the algorithms will learn to anticipate the strategic behavior of informed agents without observing it in the training samples (see a theoretical analysis in Bjorkegren, Blumenstock, and Knight, 2020).

[8] In their classical work, Goodhart's (1975) Law and Lucas (1976) Critique generalize the phenomenon in the setting of macro policy interventions.

[9] LM (2011) acknowledged the theoretical possibility that "[k]nowing that readers are using a document to evaluate the value of a firm, writers are likely to be circumspect and avoid negative language" without providing evidence.

II. Data, Variable Construction, and Sample Overview

A. Data sources

The primary data source of this study is the Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system and the associated Log File Data Set. Since 1994, the SEC has provided the public with access to securities filings containing value-relevant and market-moving information through its EDGAR system, available through the SEC's website and WRDS SEC Analytics Suite.

While EDGAR is a content archive, its Log File tracks the traffic of requests and downloads. More specifically, it comprises all records of the requests of SEC filings on EDGAR system since January 2003. Each observation in the original dataset contains information on the visitor's Internet Protocol (IP) address, timestamp, and the unique accession number of the filing that the visitor downloads. In pre-processing the raw Log File, we exclude requests that land on index pages because such requests do not download actual company filings. We then match the accession number with the SEC master filing index to select all the 10-K and 10-Q filings.[10] This procedure yields a total of 438,752 filings (119,135 10-K and 319,617 10-Q). After matching to CRSP/Compustat, our final sample of raw filings consists of 359,819 filings (90,437 10-K and 269,382 10-Q), filed by 13,763 unique CIKs, between 2003 and 2016.

Needless to say, regulatory filings are one of the venues through which firms can communicate to the marketplace. Alternatively, firms can host corporate events such as

[10] We do not include amendments and other variant filings because these documents likely mirror the original filings.

conference calls, corporate presentations, and non-deal roadshows. Regulatory filings have the advantage that the composition of the audience is mostly exogenous to firms' own decisions, which is less true in the other settings. For example, managers can invite selected audience in corporate events, while regulatory filings are open to everyone (Cohen, Lou, and Malloy, 2019). For this reason, this paper focuses on the two most important SEC filings for public companies.

B. Construction of main variables

B1. Machine Downloads

Several constructed variables are instrumental in our analyses, which we describe in detail. The first key variable measures the frequency of machine downloads of corporate filings, which serves as an upper bound as well as a proxy for the presence of "machine readers." Despite the advent of multiple data sources, the SEC EDGAR website remains the earliest and most authoritative source for company filings to be publicly released.[11] With the advances in computing power and availability of data, some large hedge funds and asset managers have started big-data driven programs to process and analyze unstructured data including corporate filings and news.[12] Recent academic studies also provide evidence that investment companies rely on machine downloads of EDGAR filings for some of their trading strategies. Crane, Crotty, and Umar (2020) find that hedge funds that employ robotic downloads perform better than those that do not. Cao,

[11] There was a multi-year episode of early leakage, which was largely resolved in mid-2015. See Bolandnazar, Jackson, Jiang, and Mitts (2020).

[12] See, e.g., "Cohen's Point72 Hires 30 People for Big Data Investing," Simone Foxman, Bloomberg, March 10, 2015, and "BlackRock Uses Big Data for Big Gains," Sarah Max, Barron's, December 26, 2015.

Du, Yang, and Zhang (2020) show that machine downloaders exhibit skills in identifying profitable copycat trades from their peers' disclosure.

To measure machine downloads, we identify an IP address downloading more than 50 unique firms' filings on any given date as a machine (i.e., robot) visitor and classify its requests on that day as machine downloads, the same criterion as used by Lee, Ma, and Wang (2015).[13] In addition, we include requests that are attributed to web crawlers in the SEC Log File Data as machine-initiated. All remaining requests are labeled as "other" requests. Finally, we aggregate machine requests and other requests, respectively, for each filing within seven days (i.e., days [0,7]) after it becomes available on EDGAR.[14]

Figure 1 shows the exponential growth of machine downloads since 2003. The number of machine downloads of corporate 10-K and 10-Q filings increased from 360,861 in 2003 to 165,318,719 in 2016.[15] During the same period, machine downloads have also become the predominant force among all EDGAR requests: the number of machine downloads as a fraction of all downloads increased from 39% in 2003 to 78% in 2016.

[Insert Figure 1 here.]

The variable *Machine Downloads* measures the *propensity* of machine downloads of a particular filing using ex ante information only. For a firm's (indexed by *i*) filing (indexed by *j*)

[13] Loughran and McDonald (2017) proposed an alternative and more aggressive approach to classify those daily IP addresses having more than 50 requests as robot visitors. Because this approach tends to classify almost all downloads as machine-driven in the most recent years, we resort to the more stringent measure by Lee, Ma, and Wang (2015). We nevertheless presented the results using the Loughran and McDonald (2017) classification, which are qualitatively similar, in sensitivity checks.

[14] We use seven days because the majority of requests happened within the first week. Results are robust under alternative cutoffs including 14 days and 30 days.

[15] There are other filings, notably 8-K, that are of strong interest to the market. We do not include 8-K filings mainly because 8-Ks, unlike 10-K/Qs, do not follow a standard structure, making it difficult to compare readability and writing styles in the cross section.

on day *t*, *Machine Downloads* is the natural logarithm of the average number of machine downloads of firm *i*'s historical filings that were filed during days [*t* – 390, *t* – 30] (we only include the machine downloads of a historical filing within seven days of posting on EDGAR, as explained earlier). *Other Downloads* (the remainder) and *Total Downloads* (the sum) are constructed analogously. Further, *%Machine Downloads* is defined as the ratio of *Machine Downloads* to *Total Downloads.*

B2. Machine Readability

The second key variable pertains to the "machine readability" of a 10-K or 10-Q filing, which measures the ease at which a filing can be "understood," i.e., processed and parsed, by an automated program. Recent literature in Accounting and Finance has studied various concepts (e.g., Hodge, Kennedy, and Maines, 2004, Blankespoor, 2019, Blankespoor, deHaan, and Marinovic, 2020, Gao and Huang, 2020) and proposed metrics (Allee, DeAngelis and Moon, 2018) of information processing costs, related to either machine or human processing costs (or both). After reviewing the existing research, especially Allee, DeAngelis and Moon (2018), we summarize the most important attributes distinctly related to machine readability as follows:[16] (i) *Table Extraction*, the ease of separating tables from text; (ii) *Number Extraction*, the ease of extracting numbers from text; (iii) *Table Format*, the ease of identifying the information contained in the table (e.g., whether a table has headings, column headings, row separators, and cell separators); (iv) *Self-Containedness*, whether a filing includes all needed information (i.e., without relying on external exhibits); and (v) *Standard Characters*, the proportion of characters that are standard ASCII (American Standard Code for Information Interchange) characters. In our main

[16] We thank Robbie Moon for sharing part of the data used in the paper.

specification, each attribute is standardized to a Z-score before being averaged to form a single-index *Machine Readability*. We present sensitivity checks using the first principal component of the five attributes as well as the individual underlying attributes.

Figure 2 shows the trend of *Machine Readability* from 2004 to 2015. *Machine Readability* saw steep ascendance till 2008, followed by modest growth before leveling off around 2011. The increasing trend per se is prima facie evidence that companies are not following a fixed template for financial filings, but instead have been adapting the format of their filings to a changing environment.[17]

[Insert Figure 2 here.]

Appendix B provides some intuition behind the *Machine Readability* variable by showing two sample filings: one with a low score (-1.09, or 1.90 standard deviation below the mean) APPLEBEES INTERNATIONAL INC in 2005, and one with a high readability score (0.31, or 0.57 standard deviation above the mean) by VIASAT INC in 2012. A comparison of the two filings is revealing.

In the excerpt for the first filings, the first "table" is surrounded by text rather than enclosed with the "<Table>… </Table>" tags, making it computationally difficult to recognize the location of a "table." Next, the filing refers to more than ten external exhibits (e.g., "form10kexhf_032905.htm"), which are not included in the filing. The excerpt of the second filing,

[17] On April 13, 2009, SEC released a mandate on "Interactive Data to Improve Financial Reporting" (see https://www.sec.gov/info/smallbus/secg/interactivedata-secg.htm). This mandate applies to financial reports of all companies and was implemented over the period 2009-2011. It requires companies to provide financial statements in interactive data format using the eXtensible Business Reporting Language (XBRL). The release states that "The new rules are intended not only to make financial information easier for investors to analyze, but also to assist in automating regulatory filings and business information processing." The mandate represents a regulatory effort in adapting disclosure to the machine readers.

in contrast, contains tags such as <Table>, <TR> (tag for row), and <TD> (tag for data cell), making it an easier task for machines to identify a table, a row in the table, and a cell in the table. Furthermore, this filing does not refer to an external exhibit.

B3. (Negative) sentiment

The third class of key variables aims at measuring "sentiments," which broadly refer to the use of natural language processing, text analysis and computational linguistics to systematically identify, extract, and quantify subjective information. Because a primary interest of this study is to contrast the sentiment as perceived by human and machine readers, we resort to two established lexica that guide the classification of sentiments by the two types of readers. The first lexicon is the Harvard General Inquirer IV-4 psychological dictionary. This comprehensive dictionary assigns 77 psychological intonations or categories to English words. For each corporate filing, we count the number of words that fall into the "Negative" category and normalize it by the length of the document, which is the total number of words in the textual part of a 10-K/Q filing, with all tags, tables, and exhibits removed, following the standard procedure of processing full 10-K documents in the literature (e.g., Loughran and McDonald, 2011; and Cohen, Malloy, and Nguyen, 2020). The resulting measure, expressed in percentage points, is termed *Harvard Sentiment*. The average filing in our sample contains four Harvard General Inquirer negative words per 100 words. The second lexicon is developed by Loughran and McDonald (2011), who create dictionaries of positive and negative words that are specific to the context of financial documents. We count the number of LM negative words and scale it by the length of the document. The resulting measure, expressed in percentage points, is the *LM Sentiment*.[18] An average (median) filing uses 1.63 (1.54)

[18] We consider only the negative sentiment related to both dictionaries because the previous literature, including Tetlock (2007), LM (2011), and Cohen, Malloy, and Nguyen (2020), find that positive sentiment is not as informative. Replacing the negative sentiment measure by a net sentiment measure does not change our results qualitatively.

LM negative words in every 100 words. The interquartile range is from 1.19 to 1.98 words per 100 words. Finally, we form the difference, *LM – Harvard Sentiment*, to capture the contrast.

B4. Additional sentiment measures

The fourth class of key variables build on Loughran and McDonald (2011)'s list of measures for broader sentiment, including litigiousness, uncertainty, weak modal and strong modal words, all in financial contexts. We extend sentiment measures to these additional attributes because Loughran and McDonald (2011) find that the frequency of words falling into these categories in firm filings is associated with stock market reactions. More specifically, *Litigious* is the number of litigation-related words (such as "claimant" and "tort") divided by the length of the document, expressed in percentage points. The other measures are constructed analogously. *Uncertainty* words capture a general notion of imprecision (such as "approximate" and "contingency"), *Weak Modal* and *Strong Modal* words convey levels of confidence (such as "always" and "must" as strong, and "possibly" and "could" as weak). In an average filing, every 100 words contain 0.97 (1.43, 0.52, and 0.30) litigious (uncertainty, weak modal, and strong modal) words.

B5. Vocal emotions

Though the focus of this study rests on 10-K and 10-Q filings, we extend to conference calls between firms and the public. The last set of key variables thus concerns audio quality. We build a web-crawler using *Selenium-Python* to obtain the audios of conference calls from 2010 to

2016 from EarningsCast.[19] After matching with CRSP/Compustat, our sample consists of 43,462 audio files from 3,290 unique firms (*gvkey*).

Anecdotal evidence suggests that executives have become aware that their speech patterns and emotions, evaluated by human or software, impact their assessment by investors and analysts.[20] A pioneer academic study by Mayew and Venkatachalam (2012) finds that analysts incorporate managers' emotions during conference calls when they make stock recommendations. One of the most prominent models of emotion, the Circumplex model, originally developed by Russell (1980), suggests that emotions are distributed in a two-dimensional space defined by valence and arousal. Following Hu and Ma (2020), we rely on a pre-trained Python machine learning package *pyAudioAnalysis* [21] (Giannakopoulos, 2015) to code the vocal emotion of each conference call. *Emotion Valence* described the extent to which an emotion is positive or negative, with a larger value indicating greater positivity. *Emotion Arousal* refers to the intensity or the strength of the associated emotion state. Both measures are bounded between –1 and 1, and a greater (lower) value suggests that the speaker is more excited (calmer).

B6. Firm characteristics

As usual, the firm characteristics variables (serving as control variables) are retrieved or based on information from standard databases accessed via WRDS, such as CRSP/Compustat, and Thomson Reuters Ownership Database. In this category of variables, *Size* is the market

[19] EarningsCast is a commercial aggregator for company earnings calls, calendar feed and podcast feed. Its website is https://earningscast.com/. *Selenium-Python*, *Selenium-Python* is an open-source software package that allows us to program a specific mouse-clicking sequential pattern for a particular website so that we can automate web browsing and internet data retrieval from the website, see https://selenium-python.readthedocs.io/.

[20] Sources: "Can Executives' Speech Patterns Provide a Good Investment Guide?" Katherine Heires, Institutional Investors, March 22, 2012. "Listening Without Prejudice: How the Experts Analyze Earnings Calls for Lies, Bluffs, and Other Flags", Sterling Wong, Minyanville, April 18, 2012.

[21] The open-source *pyAudioAnalysis* is available at https://github.com/tyiannak/pyAudioAnalysis.

capitalization in logarithm. *Tobin's Q* is the natural logarithm of the ratio of the sum of market value of equity and book value of debt to the sum of book value of equity and book value of debt. *ROA* is the ratio of EBITDA to assets. *Leverage* is the ratio of total debt to assets. *Growth* is the average sales growth of the past three years. *IndAdjRet* is the monthly average SIC3-adjusted stock returns over the past year. *InstOwnership* is the ratio of the total shares of institutional ownership to shares outstanding. *Log(#analyst)* is the natural log of one plus the number of IBES analyst covering the stock. *IdioVol* is the annualized idiosyncratic volatility (using daily data) from Fama-French three factor model. *Turnover* is the monthly average of the ratio of trading volumes to shares outstanding. *Segment* is the number of business segments and measures the complexity of business operations, following Cohen and Lou (2012). All control variables are constructed annually using information available at the previous year-end. All potentially unbounded variables are winsorized at the 1% extremes.

The summary statistics are reported in Table 1. Because multiple variables require historical information, the sample for our regression analyses start in 2004 and consists of a total of 324,607 filings (81,075 10-K and 243,532 10-Q).

[Insert Table 1 here.]

III. AI Readership and Machine Readability of Corporate Disclosure

A. Determinants of machine downloads

Since *Machine Download* is a key variable in our analysis, we first try to understand what factors drive its variation. For this purpose, we estimate a regression with *Machine Downloads* (or % *Machine Downloads)* as the dependent variable, and various firm characteristics as

independent variables. Table 2 reports the regressions all of which include year fixed effects, as well as progressively industry and firm fixed effects. Unless otherwise stated, we use the 5% level as criterion for statistical significance.

[Insert Table 2 here.]

We find that in the cross section (columns (1) and (2)), firms with larger size, lower valuation (*Tobin's Q)*, higher leverage, lower asset growth, more segments, high trading turnover, higher idiosyncratic volatility, and firms that underperform their peers in the same industry tend to have greater *Machine Downloads.* This suggests that machines tend to download filings from more mature firms with more firm-specific developments. These relations pertain to within-firm (column (3)), except that firms attract more machine downloads when it exhibits lower than usual idiosyncratic volatility and have lower analyst coverage.

Using *% Machine Downloads* reaches the opposite inferences regarding *Size, Leverage, IndAdjRet, Turnover, and Segment.* That is, small firms with high recent returns have a higher concentration of machine downloads. The contrast highlights the different determinants for the scale and concentration of machine downloads. Because our research question concerns the consequence of machine readership, the magnitude of machine downloads is the more pertinent metric and hence *Machine Downloads* is our default measure. Moreover, to the extent that firm characteristics are correlated with machine downloads, we include this list of variables in future regressions as controls.

B. Relation between machine downloads and machine readability of reports

As more and more investors use AI tools such as natural language processing and sentiment analyses, we hypothesize that companies adjust the way they talk in order to communicate effectively and predictably. A diagnostic test is thus to relate *Machine Readability* to *Machine Downloads* in the cross section and over time. Table 3 reports the results from the following regression at the filing level, indexed by firm(*i*)-filing(*j*)-date(*t*), with both year and firm (or industry) fixed effects:

$$Machine\ Readability_{i,j,t} = \beta Machine\ Downloads_{i,j,t} + \delta Other\ Downloads_{i,j,t} + \boldsymbol{\gamma}\ Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \varepsilon_{i,j,t}. \quad (1)$$

[Insert Table 3 here.]

Table 3 Panel A shows that higher machine downloads expected for a filing of a company, whether measured as the volume or percentage of machine downloads, significantly (at the 1% level) predicts more machine-reading friendly reports across all specifications. The first four columns show that a one-standard deviation increase in *Machine Downloads* is associated with 0.18 to 0.24 standard deviation increase in *Machine Readability*. The effects are almost invariant with or without the control variables, indicating that other firm characteristics have little confounding effect. The last two columns show that % *Machine Downloads* bears a very similar relation to machine readability, where a one-standard deviation increase in % *Machine Downloads* predicts a 0.04 to 0.05 standard deviation increase in *Machine Readability*.

For the hypothesis that firms accommodate machine readers to be supported it is equally important that the data show an absence of correlation between *Machine Readability* and *Other Downloads*. That is, the other, presumably non-machine downloads serve as a natural placebo test.

Indeed, all four coefficients on *Other Downloads* (columns (1) to (4)) turn out to be indistinguishable from zero, economically and statistically.

Panel B of Table 3 presents results from specifications using alternative definitions of *Machine Readability*. In the first two columns, the dependent variable is the principal component of the five attributes characterizing machine readability. The last two columns of Panel B adopt the Loughran and McDonald (2017) definition of machine downloads, which classifies more downloads as machine-driven. All four specifications show that *Machine Downloads* is significantly (at the 1% level) associated with, but *Other Downloads* exhibits no positive relation with, *Machine Readability*. In fact, higher *Other Downloads* is negatively and significantly associated with machine-friendly format in reporting.

Panel C of Table 3 breaks down *Machine Readability* into its five components: *Table Extraction*, *Number Extraction*, *Table Format*, *Self-Containedness*, and *Standard Characters*. Results show that high expected machine downloads increase all five sub-metrics of machine readability significantly (at the 1% level). Again, the coefficients of *Other Downloads* do not have consistent signs across the five attributes.

C. Cross validation of Machine Downloads and Machine Readability as empirical proxies

Our analyses to follow critically depend on *Machine Downloads* and *Machine Readability* being effective proxies for the presence of machine readership and the ease with which machine can process the filings. We thus conduct multiple tests that support the validity of the two key empirical proxies. First, we connect *Machine Downloads* to its primary suspect, hedge funds who adopt AI strategies. Following Guo and Shi (2020), we classify a hedge fund to be AI-prone if

there is at least one employee who has been involved in AI projects based on their LinkedIn profiles.[22] We then define *AI Hedge Fund* to be the percentage of shares outstanding that is held by such hedge funds at the firm-quarter level, based on the 13F filings via Thomson Reuters Ownership database. The last two columns of Table 2 include *AI Hedge Fund* as an additional variable to predict *Machine Downloads.* Indeed, the coefficients are positive and significant.

Second, we conjecture and test that machine readers should lead to faster trades after a filing is posted, given their natural advantage in the capacity and speed processing information. Moreover, such an advantage should be elevated when the files are composed to be machine friendly. Such a hypothesis is operationalized into a duration analysis connecting "time to trade" and the key independent variables. Using high-frequency data, we conduct the following regression at the filing level, indexed by firm(*i*)-filing(*j*)-date(*t*), with year and firm (or industry) fixed effects:

$$Time\ to\ Trade_{i,j,t} = \beta_1 Machine\ Downloads_{i,j,t} \times Machine\ Readability_{i,j,t} + \beta_2 Machine\ Downloads_{i,j,t} + \beta_3 Machine\ Readability_{i,j,t} + \delta Other\ Downloads_{i,j,t} + \boldsymbol{\gamma}\ Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \varepsilon_{i,t}. \quad (2)$$

There are two versions for the dependent variable: *Time to the First Trade* and *Time to the First Directional Trade,* the construction of which follow Bolandnazar, Jackson, Jiang, and Mitts (2020). *Time to the First Trade* is the length of time, in seconds, between the time stamps of EDGAR posting and the first trade of the issuer's stock afterwards. *Time to the First Directional Trade* adds a requirement that the trade needs to be profitable (before any transaction cost) based on the price at the end of the 15th minute post filing. That is, the first directional trade is the first

[22] We thank Xuxi Guo and Zhen Shi for sharing the data of hedge funds with AI-experienced employees. AI projects are identified based on both job title and descriptions of experience/responsibility.

buy (sell) trade at a price below (above) the terminal value, where buy- and sell-initiated trades are classified by the Lee and Ready (1991) algorithm. As in Bolandnazar et al. (2020), we focus on the 15-minute window in order to isolate the effect of the filing; and hence both variables are censored at the end of the time window.

The results, reported in Table 4, support the prediction that high *Machine Downloads* are associated with faster trades after a filing becomes publicly available. A one-standard deviation increase in *Machine Downloads* saves 8.56 to 14.73 seconds for the first trade and 13.29 to 21.80 seconds for the first directional trade. All coefficients associated with directional trades (in the last four columns) are significant at the 1% level, while the coefficients lose significance with *Time to the First Trade* when firm fixed effects are included. Moreover, the relation between *Machine Downloads* and *Time to Trade* is indeed significantly stronger when *Machine Readability* is higher. The test, in addition to serving as a joint validation of the two key empirical proxies, but also demonstrates the real impact of AI adoption in trading. To the extent that faster market reaction to corporate disclosure is a sign of effective communication of firms' financial health, which results in efficient information dissemination, the result also justifies the need to firms to cater to the machines.

[Insert Table 4 here.]

IV. Managing Sentiment and Tones with Machine Readers

A. Textual sentiment

While truthfulness in disclosure reports is expected and required, it is well known that corporate disclosures are usually positively biased without necessarily crossing the line of law and

compliance (Loughran and McDonald, 2011; Kothari, Shu, and Wysocki, 2009). Company executives usually want to portray their business activity in the most positive light to attract or gain from stakeholders (creditors, employees, suppliers, and customers). The messages from the top management may also include the sentiment that these people truly believe but are based on unintentional misconceptions, which, though, still end up being self-serving.

For this reason, the prior literature has documented that "positive" words are far more common than "negative" words in corporate reports, based on respectable lexicons such as the Harvard Psychosociological Dictionary, specifically, the Harvard-IV-4 TagNeg (H4N) file. Such a list of words were originally developed for human readers and for general purposes, and over time they serve as an objective standard for researchers to analyze the sources and consequences of tones and sentiments in corporate disclosures and new media as perceived by the general readership (Tetlock, 2007, Tetlock, Saar-Tsechansky, and Macskassy, 2008, Hanley and Hoberg, 2010).

However, the meaning and tone of English words are highly context and discipline specific, and a general word categorization scheme might not translate effectively into a specialized field such as finance. This motivated the influential work by Loughran and McDonald (2011), which presented a specialized dictionary of positive and negative words that fits the unique text of financial situations. In fact, according to Loughran and McDonald (2011), almost three-fourth of the words identified by the Harvard Dictionary as negative (such as "liability") are words typically not considered negative in financial contexts. The dictionary has since become the leading lexicon used in algorithms for sentiment calibration.[23]

[23] For example, as of May 2020, the LM paper has been cited more than 2,300 times by researchers. And their word list has been adopted by the WRDS SEC Sentiment Data. The dictionary has been frequently featured in industry

The timeline of Harvard General Inquirer dictionary (existed since 1996) and the Loughran-McDonald dictionary (since 2011)[24] and their differential adoption by human versus machine readers, provide a unique setting for us to test how the writing of corporate filings adjusts to AI readers. We consider the following regression at the filing level, indexed by firm(*i*)-filing(*j*)-date(*t*), with year and firm (or industry) fixed effects:

$$Negative\ Sentiment_{i,j,t} = \beta_1 Machine\ Downloads_{i,j,t} \times Post_t + \beta_2 Machine\ Downloads_{i,j,t} + \delta Other\ Downloads_{i,j,t} + \boldsymbol{\gamma}\ Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \varepsilon_{i,t}. \quad (3)$$

There are three versions of the dependent variable *Negative Sentiment* in the equation above: the *LM Sentiment*, the *Harvard Sentiment*, and their difference *LM – Harvard Sentiment*, as defined in Section II.B3. We consider the prevalence of negative words only because earlier research (Tetlock, 2007; Loughran and McDonald, 2011; Cohen, Lou, and Malloy, 2020) indicate that positive words are not informative of firm future outcomes or stock returns. *Post* is an indicator variable for years that came after the publication of Loughran and McDonald (2011), which is equal to one for filings in 2012 and onwards, and zero otherwise. Filings in 2011 are excluded from the analysis. The year fixed effect subsumes the variable *Post* on its own.

Under the hypothesis that AI readers employed by algorithmic investors shape the style and quality of corporate writing, we expect coefficient β_1 to be significantly negative for *LM Sentiment* but not for *Harvard Sentiment*, and the relation between *LM Sentiment* and *Machine*

white papers and technical reports, such as in "Natural Language Processing in Finance: Shakespeare Without the Monkeys" by the Man Group in July 2019.

[24] The paper was in public distribution, e.g., posted on the SSRN, since 2009. Google citation counts show that Loughran and McDonald (2011) was cited 10 times prior to 2011, 243 times by 2013, and has grown exponentially to 2,483 times as of September 2020.

Downloads should be present primarily during the *Post* period (after the publication of Loughran and McDonald (2011)). Such an exclusive set of effects is confirmed by results in Table 5.

[Insert Table 5 here.]

Table 5 shows an unambiguous contrast before and after 2011 on the effect of measures related to Loughran and McDonald (2011), the year when the paper was published. Post 2011, a one-standard deviation increase in *Machine Downloads* is associated with a 9-11 basis points incremental decrease in *LM Sentiment*, on top of an insignificant (column (3) with industry fixed effect) or much smaller (column (4) with firm fixed effects) effect during the pre-2011 period. The incremental effect post-2011 represents about 5% of the sample mean of *LM Sentiment*, or 0.15 standard deviations, and is significant at the 1% level. In contrast, *Harvard Sentiment* does not bear any negative relation with *Machine Downloads* (columns (5) and (6)). Finally, columns (1) and (2) show that the relation between *LM – Harvard Sentiment* and *Machine Downloads* conforms to that of *LM Sentiment*, confirming that the differential effect is mainly driven by reduced *LM Sentiment.*

Results in Table 5 keep the possibility open that the publication of Loughran and McDonald (2011) merely reflects a general trend of a strengthening relation between the machine downloads and avoiding using words that are perceived to have negative annotations in the finance context. Such a possibility still supports the general thesis that machine readership impacts disclosure quality; nevertheless, a "parallel pre-trend" would allow a sharper identification on the impact of a new lexicon available to machine reading. Figure 3 illustrate the structural break, instead of a pre-existing and continuing trend, around 2011. More specifically, we aggregate the *LM – Harvard Sentiment* at the annual level, separately for filings that are in the top and bottom terciles of *Machine Downloads* in each year. Figure 3 Panel A plots the time series of the

incremental tendency to use LM-negative words (for finance context) over Harvard-negative words (for general context) by the two groups of filings.

[Insert Figure 3 here.]

Panel A of Figure 3 shows a parallel pre-trend of the two groups till 2011 and then a clear divergence afterward. Before 2011, filings in the top and bottom terciles of *Machine Downloads* exhibit clustered movements in the *LM – Harvard Sentiment*. Afterwards, the sentiment of the top tercile trends down relative to that of the bottom tercile. Panel B of Figure 3 takes a different sorting method by separating filings into the top quartile of *Machine Downloads* from the rest. The resulting graph confirms the parallel pre-trend and then divergence around 2011, suggesting that disclosures with the highest expected machine readership are driving the results.

Given the quasi-randomness of the exact year of publication, it is unlikely that the publication of Loughran and McDonald (2011) made the perfect timing on the structural break. In other words, it is implausible that the LM dictionary summarizes the practice that was already in place, and that it serves as an exactly timed concurrent side-show. Table 5 and Figure 3 thus support the hypothesis that corporate writing has been adjusted to serve machine readers, which was impacted by the availability of the LM dictionary.

B. Managing other textual tones with machine readers

In addition to providing lists of sentimental words, Loughran and McDonald (2011) also construct lists of "tone" words aiming to capture litigiousness, uncertainty, and weak and strong modality that are tailored to the financial context. The expanded dictionary allows machines to assess more dimensions of the annotations of a document. Loughran and McDonald (2011)

discover that stock market respond more positively to disclosure using fewer negative, uncertain, modal strong, and modal weak words; and that firms with a high proportion of negative or strong modal words are more likely to report material weakness. Given the market perception, it is curious to see whether managers also adjust tones along these dimensions after the methodology became publicly known. We re-estimate Equation (3) by replacing the dependent variable by *Litigious*, *Uncertainty*, *Weak Modal*, and *Strong Modal*, which are all defined in Section II.B4 as well as in Appendix A:

$$Tone_{i,j,t} = \beta_1 Machine\ Downloads_{i,j,t} \times Post_t + \beta_2 Machine\ Downloads_{i,j,t} + \delta Other\ Downloads_{i,j,t} + \boldsymbol{\gamma}\, Control_{i,year} + \alpha_i (or\ \alpha_{SIC3}) + \alpha_{year} + \varepsilon_{i,t}. \quad (4)$$

To the extent that managers have adjusted the frequency of LM-negative words based on their knowledge about investor reaction to sentiment they should, then, also understand the impact of other tones documented in Loughran and McDonald (2011). Given Loughran and McDonald's (2011) discovery that the frequency of all four tones were met with negative stock market reactions, we conjecture that managers of firms with high expected machine readership should tone down these words after 2011. Results in Table 6 support such a prediction. The coefficients associated with *Machine Downloads* × *Post* are significant (at 5% level or less) for all four dependent variables. That is, post-2011 corporate reports that are expected to be read by machines avoid convey a sentiment, which could come out of an algorithm, that is predictive of legal liabilities, that is indicative of uncertain prospects, and that exhibit too little or too much confidence and surety. For example, a one-standard deviation increase in *Machine Downloads* predicts a 0.19 standard deviation decrease in the *Litigious* tone.

[Insert Table 6 here.]

C. Managing audio quality in conference calls with machine readers

Though the textual quality of disclosures is the focus of this study, voice analytics, enabled by the development of modern machine learning methods, provides an out-of-sample test. Starting around 2008, voice analytic software, such as the commercial Layered Voice Analysis (LVA) software and open-source software on GitHub, have gained attention among investors looking for an edge in information processing. Such software has enabled researchers to study the vocal expressions of managers and their implications on capital markets (Mayew and Ventakachalam, 2012; and Hu and Ma, 2020). If managers are aware that their disclosure documents could be parsed by machines, then they should also expect that their machine readers may also be using voice analyzers to extract signals from vocal patterns and emotions contained in managers' speeches.

This section explores whether the management adjust the way they talk (on conference call) when they expect that machines are listening, based on a sample of audio data of conference calls from 2010 to 2016, as described in Section II.B5. Since there are no data on downloads of conference calls, we keep *Machine Downloads* of a firm's filings as the proxy for the prevalence of "machine listeners," based on the premise that *Machine Downloads* represents the propensity of investors to deploy AI tools in analyzing corporate disclosure. Table 7 reports the results from the following regression at the conference call level, indexed by firm (*i*)-call (*k*)-date (*t*), with both year and firm (or industry) fixed effects:

$$Emotion_{i,k,t} = \beta Machine\ Downloads_{i,k,t} + \delta Other\ Downloads_{i,k,t} + \boldsymbol{\gamma}\ Control_{i,year} + \alpha_i(\alpha_{SIC3}) + \alpha_{year} + \varepsilon_{i,k,t}. \quad (5)$$

We measure emotion along two dimensions developed in psychology, *Valence* and *Arousal,* that captures and positivity and intensity of vocal tones (Russell, 1980).

[Insert Table 7 here.]

The first four columns of Table 7 show that higher *Machine Downloads* is associated with higher *Valence,* or positivity in vocal emotion. A one-standard deviation increase in *Machine Downloads* is associated with a 0.28 standard deviation higher *Valence.* Last four columns of Table 7 indicate a positive, but much weaker, relation between *Machine Downloads* and *Arousal,* i.e., a more exciting emotion in conference calls. Note that tone of positivity or excitement could be driven by the fundamentals, the health of the earnings in this case. For this reason, we further include *EarningsSurprise*, defined as the difference between actual earnings and median analyst forecast, in columns (4) and (8) as an additional control variable.[25] The coefficients associated with *Machine Downloads* barely change.

Based on videos of entrepreneurs pitching investors for funding, Hu and Ma (2020) show that venture capitalists are more likely to invest in start-ups whose founders give pitches that are rated high in valence and arousal. It is plausible that reactions of VC investors to vocal emotion also apply to the general capital markets. Our findings support the hypothesis that managers may manipulate their vocal expressions to achieve a more favorable effect on investors that rely on machine processing, and also justifies the anecdotal evidence that managers increasingly seek professional coaching in order to improve vocal performances.[26]

[25] Calculating the *EarningsSurprise* variable requires analyst coverage (tracked by the I/B/E/S analyst data), which results in a much smaller sample.

[26] Sources: "Listening Without Prejudice: How the Experts Analyze Earnings Calls for Lies, Bluffs, and Other Flags", Sterling Wong, Minyanville, April 18, 2012. "How to listen for the hidden data in earnings calls", Alina Dizik, Chicago Booth Review, May 25, 2017.

IV. Concluding Remarks

This paper presents the first study showing how corporate disclosure in writing and speaking has been reshaped by machine readership employed by algorithmic traders and quantitative analysts. Our findings indicate that increasing AI readership motivates firms to prepare filings that are more friendly to machine parsing and processing, highlighting the growing roles of AI in the financial markets and their potential impact on corporate decisions. Firms manage sentiment and tone perception that is catered to AI readers by differentially avoiding words that are perceived as negative by algorithms, as compared to those by human readers. While the literature has shown how investors and researchers apply machine learning and computational tools to extract information from disclosure and news, our study is the first to identify and analyze the *feedback effect*, i.e., how companies adjust the way they talk knowing that machines are listening. Such a feedback effect can lead to unexpected outcomes, such as manipulation and collusion (Calvano, Calzolari, Denicolo, and Pastorello, 2019). The technology advancement calls for more studies to understand the impact of and induced behavior by AI in financial economics.

Appendix A: Definitions of Variables

Variable	Definition
Machine Downloads	For a firm's filing on day *t*, *Machine Downloads* is the natural logarithm of the average number of machine downloads of the firm's historical filings that were filed during days [*t* – 390, *t* – 30] days. To measure machine downloads, we identify an IP address downloading more than 50 unique firms' filings daily as a machine (i.e., robot) visitor, the same criterion as used by Lee, Ma, and Wang (2015). In addition, we include requests that are attributed to web crawlers in the SEC Log File Data as machine-initiated. Machine requests are aggregated for each filing within seven days (i.e., days [0,7]) after it becomes available on EDGAR.
Other Downloads	For a firm's filing on day *t*, *Other Downloads* is the natural logarithm of the average number of non-machine downloads of the firm's historical filings that were filed during days [*t* – 390, *t* – 30] days.
Total Downloads	For a firm's filing on day *t*, *Total Downloads* is the natural logarithm of the average number of total downloads of the firm's historical filings that were filed during days [*t* – 390, *t* – 30] days.
%Machine Downloads	*Machine Downloads/Total Downloads*
Machine Readability	*Machine Readability* is the average of five filing attributes, including (i) *Table Extraction*, the ease of separating tables from text; (ii) *Number Extraction*, the ease of extracting numbers from text; (iii) *Table Format*, the ease of identifying the information contained in the table (e.g., whether a table has headings, column headings, row separators, and cell separators); (iv) *Self-Containedness*, whether a filing includes all needed information (i.e., without relying on external exhibits); and (v) *Standard Characters*, the proportion of characters that are standard ASCII (American Standard Code for Information Interchange) characters. In our main specification, each attribute is standardized to a Z-score before being averaged to form a single-index *Machine Readability*.
PCA Machine Readability	*PCA Machine Readability* is the first principal component of the five underlying filing attributes from *Machine Readability.*
Time to the First Trade	*Time to the First Trade* is the length of time, in seconds, between the EDGAR publication time stamp and the first trade of the issuer's stock, censored at the end of a 15-minute window.
Time to the First Directional Trade	*Time to the First Directional Trade* is the length of time, in seconds, between the EDGAR publication time stamp and the first directional trade after a filing is publicly released, and it is censored at the end of the 15-minute window. The first directional trade is the first buy (sell) trade at a price below (above) the terminal value at the end of the window, where buy- and sell-initiated trades are classified by the Lee and Ready (1991) algorithm.
LM Sentiment	The number of Loughran-McDonald (LM) finance-related negative words in a filing divided by the total number of words in the filing, expressed in percentage points.
Harvard Sentiment	The number of Harvard General Inquirer negative words in a filing divided by the total number of words in the filing, expressed in percentage points.
LM – Harvard Sentiment	*LM Sentiment* minus *Harvard Sentiment.*

Litigious	The number of Loughran-McDonald (LM) litigation-related words in a filing divided by the total number of words in the filing, expressed in percentage points.
Uncertainty	The number of Loughran-McDonald (LM) uncertainty-related words in a filing divided by the total number of words in the filing, expressed in percentage points.
Weak Modal	The number of Loughran-McDonald (LM) weak modal words in a filing divided by the total number of words in the filing, expressed in percentage points.
Strong Modal	The number of Loughran-McDonald (LM) strong modal words in a filing divided by the total number of words in the filing, expressed in percentage points.
Post	*Post* is an indicator variable equal to one for filings in 2012 and onwards, and zero for filings in 2010 and before (filings in 2011 are excluded from the analysis).
Emotion-Valence	The positivity of speech emotion, calculated from a pre-trained Python machine learning package *pyAudioAnalysis.*
Emotion-Arousal	The excitedness of speech emotion, calculated from a pre-trained Python machine learning package *pyAudioAnalysis.*
Size	The natural logarithm of the market capitalization.
Tobin's Q	The natural logarithm of the ratio of the sum of market value of equity and book value of debt to the sum of book value of equity and book value of debt.
ROA	The ratio of EBITDA to assets
Leverage	The ratio of total debt to assets.
Growth	The average sales growth of the past three years.
IndAdjRet	The monthly average SIC3-adjusted stock returns over the past year.
InstOwnership	The ratio of the total shares of institutional ownership to shares outstanding.
AIHedgeFund	The percentage of shares outstanding owned by AI hedge funds, classified based on employees work experience in AI-related projects disclosed on their LinkedIn profiles.
Log(#analyst)	The natural log of one plus the number of IBES analyst covering the stock
IdioVol	The annualized idiosyncratic volatility (using daily data) from Fama-French three factor model.
Turnover	The monthly average of the ratio of trading volumes to shares outstanding.
Segment	The number of business segments and measures the complexity of business operations, following Cohen and Lou (2012).
EarningSurprise	The difference between the actual quarterly earnings and the median earnings forecast of IBES analysts scaled by price.

Appendix B. Excerpts of Two 10-K Filings

This figure shows two sample fillings, one with a low *Machine Readability* score (-1.09, or 1.90 standard deviation below the mean) by APPLEBEES INTERNATIONAL INC in 2005 and one with a high *Machine Readability* score (0.31, or 0.57 standard deviation above the mean) by VIASAT INC in 2012. *Machine Readability* is the average of five standardized filing attributes, including (i) *Table Extraction*, the ease of separating tables from text; (ii) *Number Extraction*, the ease of extracting numbers from text; (iii) *Table Format*, the ease of identifying the information contained in the table (e.g., whether a table has headings, column headings, row separators, and cell separators); (iv) *Self-Containedness*, whether a filing includes all needed information (i.e., without relying on external exhibits); and (v) *Standard Characters*, the proportion of characters that are standard ASCII (American Standard Code for Information Interchange) characters.

Excerpt 1. APPLEBEES INTERNATIONAL INC, CIK: 0000853665, March 30, 2005

We opened 32 new company Applebee's restaurants in 2004 and anticipate opening at least 40 new company Applebee's restaurants in 2005, excluding up to eight restaurants that were closed in 2004 by a former franchisee which we may re-open in Memphis, Tennessee. The following table shows the areas where our company restaurants were located as of December 26, 2004:

Area	
New England (includes Maine, Massachusetts, New Hampshire, New York, Rhode Island and Vermont).....	65
Detroit/Southern Michigan.....	62
Minneapolis/St. Paul, Minnesota.....	58
St. Louis, Missouri/Illinois.....	47
North/Central Texas.....	45
Virginia.....	42
Kansas City, Missouri/Kansas.....	33
Washington, D.C. (Maryland, Virginia).....	29
San Diego/Southern California.....	20
Las Vegas/Reno, Nevada.....	15
Albuquerque, New Mexico.....	8
	424

(omitted)

<TYPE>EX-10
<SEQUENCE>4
<FILENAME>form10kexhf_032905.htm
<DESCRIPTION>EXHIBIT 10.2
<TEXT>
<HTML>
<HEAD>
<TITLE>Exhibit 10.2</TITLE>

Excerpt 2. VIASAT INC, CIK: 0000797721, May 25, 2012

Text format for machine processing:

```
 <P STYLE="margin-top:0px;margin-bottom:0px; margin-left:2%"><FONT STYLE="font-family:Times New Roman" SIZE="2"><B><I>Amortization of acquired intangible assets </I></B>
</FONT></P>
<P STYLE="margin-top:6px;margin-bottom:0px; text-indent:4%"><FONT STYLE="font-family:Times New Roman" SIZE="2">We amortize our acquired intangible assets from prior
acquisitions over their estimated useful lives ranging from eight months to ten
years. The decrease in amortization of acquired intangible assets of approximately $0.7 million in fiscal year 2012 compared to last fiscal year was a result of an
approximately $1.2 million decrease in amortization as certain acquired technology
intangibles in our government systems and commercial networks segment became fully amortized over the preceding twelve months, offset by an increase in amortization of
approximately $0.6 million due to our acquisition of Stonewood in July 2010.
Expected amortization expense for acquired intangible assets for each of the following periods is as follows: </FONT></P> <P STYLE="font-size:12px;margin-top:0px;margin-
bottom:0px"> </P>
<TABLE CELLSPACING="0" CELLPADDING="0" WIDTH="68%" BORDER="0" STYLE="BORDER-COLLAPSE:COLLAPSE" ALIGN="center">


<TR>
<TD WIDTH="84%"></TD>
<TD VALIGN="bottom" WIDTH="10%"></TD>
<TD></TD>
<TD></TD>
<TD></TD></TR>
<TR>
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD>
<TD VALIGN="bottom"><FONT SIZE="1">  </FONT></TD>
<TD VALIGN="bottom" COLSPAN="2" ALIGN="center" STYLE="border-bottom:1px solid #000000"><FONT STYLE="font-family:Times New Roman" SIZE="1"><B>Amortization</B></FONT></TD>
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD></TR>
<TR>
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD>
<TD VALIGN="bottom"><FONT SIZE="1">  </FONT></TD>
<TD VALIGN="bottom" COLSPAN="2" ALIGN="center"><FONT STYLE="font-family:Times New Roman" SIZE="1"><B>(In thousands)</B></FONT></TD>
<TD VALIGN="bottom"><FONT SIZE="1"> </FONT></TD></TR>


<TR BGCOLOR="#cceeff">
<TD VALIGN="top"> <P STYLE="margin-left:1.00em; text-indent:-1.00em"><FONT STYLE="font-family:Times New Roman" SIZE="2">Expected for fiscal year 2013</FONT></P></TD>
<TD VALIGN="bottom"><FONT SIZE="1">  </FONT></TD>
<TD VALIGN="bottom"><FONT STYLE="font-family:Times New Roman" SIZE="2">$</FONT></TD>
<TD VALIGN="bottom" ALIGN="right"><FONT STYLE="font-family:Times New Roman" SIZE="2">15,592</FONT></TD>
<TD NOWRAP VALIGN="bottom"><FONT STYLE="font-family:Times New Roman" SIZE="2">  </FONT></TD></TR>
```

HTML as in a web browser (for the reader's convenience, the following picture shows the contents of the above scripts if shown as an HTML in a web browser):

Table of Contents

Amortization of acquired intangible assets

We amortize our acquired intangible assets from prior acquisitions over their estimated useful lives ranging from eight months to ten years. The decrease in amortization of acquired intangible assets of approximately $0.7 million in fiscal year 2012 compared to last fiscal year was a result of an approximately $1.2 million decrease in amortization as certain acquired technology intangibles in our government systems and commercial networks segment became fully amortized over the preceding twelve months, offset by an increase in amortization of approximately $0.6 million due to our acquisition of Stonewood in July 2010. Expected amortization expense for acquired intangible assets for each of the following periods is as follows:

	Amortization (In thousands)
Expected for fiscal year 2013	$ 15,592
Expected for fiscal year 2014	13,848
Expected for fiscal year 2015	13,772
Expected for fiscal year 2016	10,193
Expected for fiscal year 2017	4,626
Thereafter	5,010
	$ 63,041

References

Ahern, Kenneth R., and Denis Sosyura, 2014, Who writes the news? Corporate press releases during merger negotiations, *Journal of Finance* 69, 241–291.

Allee, Kristian D., Matthew D. DeAngelis, and James R. Moon Jr, 2018, Disclosure "scriptability", *Journal of Accounting Research* 56, 363–430.

Bernard, Darren, Terrence Blackburne, and Jacob Thornock, 2020, Information flows among rivals and corporate investment, *Journal of Financial Economics* 136, 760–779.

Björkegren, Daniel, Joshua E. Blumenstock, and Samsun Knight, 2020, Manipulation-Proof Machine Learning, Working paper, Brown University and U.C. Berkeley.

Blankespoor, Elizabeth, 2019, The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate, *Journal of Accounting Research* 57, 919–967.

Blankespoor, Elizabeth, Ed deHaan, and Ivan Marinovic, 2020, Disclosure processing costs, investors' information choice, and equity market outcomes: A review, *Journal of Accounting and Economics* forthcoming.

Bolandnazar, Mohammadreza, Robert J. Jackson Jr, Wei Jiang, and Joshua Mitts, 2020, Trading against the random expiration of private information: A natural experiment, *Journal of Finance* 75, 5–44.

Bond, Philip, Alex Edmans, and Itay Goldstein, 2012, The real effects of financial markets, *Annual Review of Financial Economics* 4, 339–360.

Calvano, Calzolari, Denicolo, and Pastorello, 2019, Artificial intelligence, algorithm pricing and collusion, *American Economic Review* forthcoming

Cao, Sean, Kai Du, Baozhong Yang, and Alan L. Zhang, 2020, Copycat skills and disclosure costs: Evidence from peer companies' digital footprints, Working paper, Georgia State University and Pennsylvania State University.

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

Chen, Mark A., Qinxi Wu, and Baozhong Yang, 2019, How valuable is FinTech innovation? *Review of Financial Studies* 32, 2062–2106.

Cohen, Lauren, and Dong Lou, 2012, Complicated firms, *Journal of Financial Economics* 104, 383–400.

Cohen, Lauren, Dong Lou, and Christopher Malloy, 2019, Playing favorites: How firms prevent the revelation of bad news, *Management Science* forthcoming.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *Journal of Finance* forthcoming.

Cong, Lin William, Tengyuan Liang, Baozhong Yang, and Xiao Zhang, 2020, Analyzing textual information at scale, *Information to Facilitate Efficient Decision Making: Big Data, Blockchain and Relevance* (ed. Kashi Balachandran), World Scientific Publishers, forthcoming

Cong, Lin William, Tengyuan Liang, and Xiao Zhang, 2019, Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information, Working paper, Chicago University and Cornell University.

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

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2011, In search of attention, *Journal of Finance* 66, 1461–1499.

Dong, Jinshuo, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu, 2018, Strategic classification from revealed preferences, *Proceedings of the 2018 ACM Conference on Economics and Computation*, 55–70.

Gao, Meng, and Jiekun Huang, 2020, Informing the market: The effect of modern information technologies on information production, *Review of Financial Studies* 33: 1367–1411.

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

Giannakopoulos, Theodoros, 2015, pyAudioAnalysis: An open-source python library for audio signal analysis, PloS one 10, e0144610.

Goodhart, Charles, 1975, Problems of monetary management: the UK experience in papers in monetary economics, *Monetary Economics* 1.

Guo, Xuxi, and Zhen Shi, 2020, The impact of AI talents on hedge fund performance, Working paper, Georgia State University.

Hanley, Kathleen Weiss, and Gerard Hoberg, 2010, The information content of IPO prospectuses, *Review of Financial Studies* 23, 2821–2864.

Hanley, Kathleen Weiss, and Gerard Hoberg, 2019, Dynamic interpretation of emerging risks in the financial sector, *Review of Financial Studies* 32, 4543–4603.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters, 2016, Strategic classification, *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Hodge, Frank D., Jane Jollineau Kennedy, and Laureen A. Maines, 2004, Does search-facilitating technology improve the transparency of financial reporting? *The Accounting Review* 79: 687–703.

Hu, Allen, and Song Ma, 2020, Human interactions and financial investment: A video-based approach, Working paper, Yale University.

Huang, Alan Guoming, Hongping Tan, Russ Wermers, 2020, Institutional Trading around Corporate News: Evidence from Textual Analysis, *Review of Financial Studies*, Forthcoming.

Hwang, Byoung-Hyoun, and Hugh Hoikwang Kim, 2017, It pays to write well, Journal of Financial Economics 124, 373–394.

Jegadeesh, Narasimhan, and Di Wu, 2013, Word power: A new approach for content analysis, *Journal of financial economics* 110, 712–729.

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

Kothari, Sabino P., Susan Shu, and Peter D. Wysocki, 2009, Do managers withhold bad news? *Journal of Accounting Research, 47*, 241–276.

Lee, Charles MC, Paul Ma, and Charles CY Wang, 2015, Search-based peer firms: Aggregating investor perceptions through internet co-searches, *Journal of Financial Economics* 116, 410–431.

Lee, Charles MC, and Mark J. Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance* 46, 733–746.

Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan, 2020, Measuring corporate culture using machine learning, *Review of Financial Studies*, Forthcoming.

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

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

Loughran, Tim, and Bill McDonald, 2017, The use of EDGAR filings by investors, *Journal of Behavioral Finance 18*, 231–248.

Lucas, Robert E, 1976, Econometric policy evaluation: A critique, *Carnegie-Rochester Conference Series on Public Policy* 1, 19–46.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Mayew, William J., and Mohan Venkatachalam, 2012, The power of voice: Managerial affective states and future firm performance, *Journal of Finance* 67, 1–43.

Russell, James A, 1980, A circumplex model of affect, *Journal of Personality and Social Psychology* 39, 1161–1178.

Milli, Smitha, John Miller, Anca D. Dragan, and Moritz Hardt, 2019, The social cost of strategic classification, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 230–239.

Tetlock, Paul C, 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

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

Figure 1. Trend of Machine Downloads

This figure plots the annual number of machine downloads (blue bars and left axis) and the annual percentage of machine downloads over total downloads (red line and right axis) across all 10-K and 10-Q filings from 2003 to 2016. Machine downloads are defined as downloads from an IP address downloading more than 50 unique firms' filings daily. The number of machine downloads and the number of total downloads for each filing are recorded as the respective downloads within seven days after the filing becomes available on EDGAR.

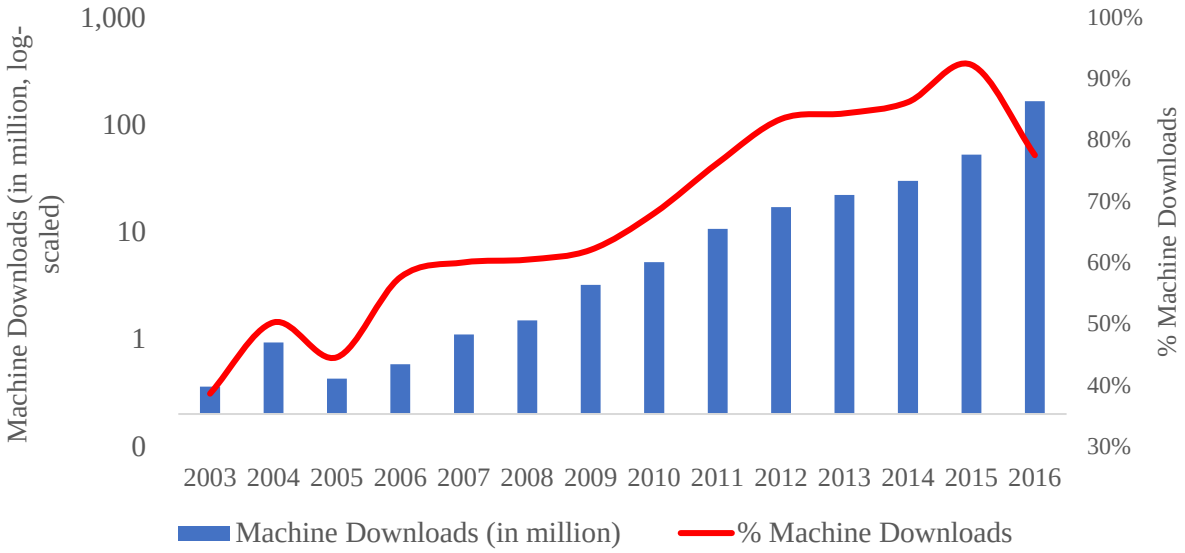


Figure 2. Trend of Machine Readability

This figure plots the annual *Machine Readability* across all 10-K and 10-Q filings from 2004 to 2015. *Machine Readability* is the average of five standardized filing attributes, including *Table Extraction*, *Number Extraction*, *Table Format*, *Self-Containedness*, and *Standard Characters*. All attributes are defined in Appendix A.

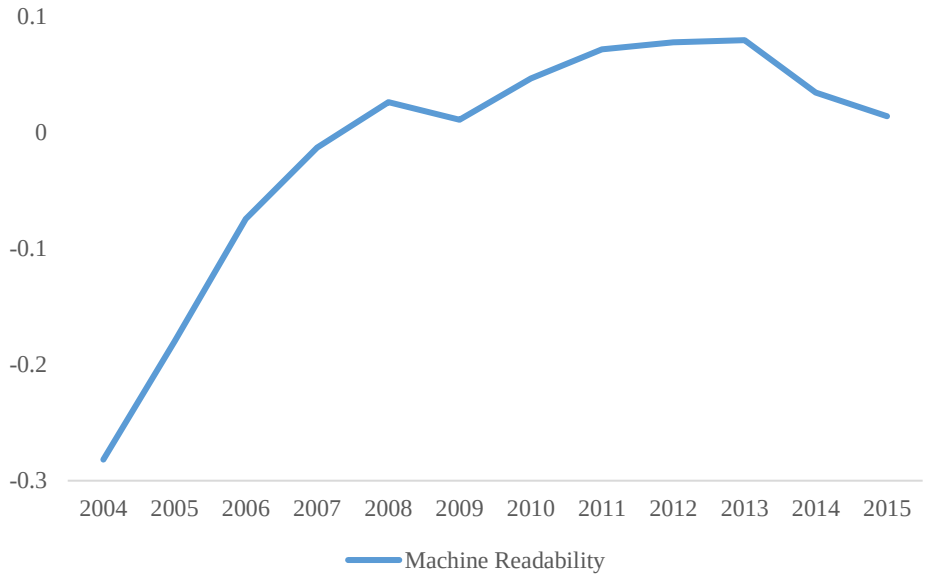
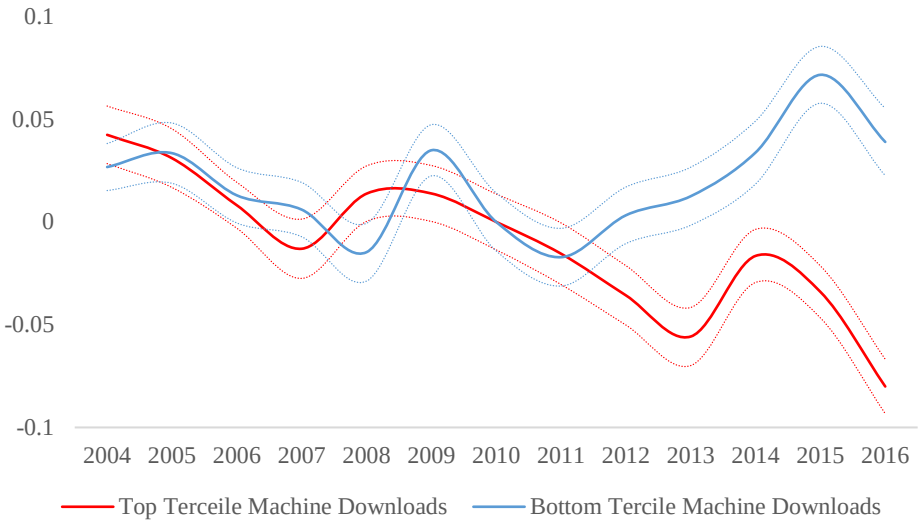


Figure 3. Sentiment Trend and Machine Downloads

This figure plots *LM – Harvard Sentiment* of 10-K and 10-Q filings and compares sentiment of firms with high machine downloads with that of the low group. *LM – Harvard Sentiment* is the difference of *LM Sentiment* and *Harvard Sentiment*. *LM Sentiment* is defined as the number of Loughran-McDonald (LM) finance-related negative words in a filing divided by the total number of words in the filing. *Harvard Sentiment* is defined as the number of Harvard General Inquirer negative words in a filing divided by the total number of words in the filing. In Panel A, filings are sorted into top tercile or bottom tercile based on *Machine Downloads*, defined in Appendix A. In Panel B, filings are sorted into top quartile or the rest based on *Machine Downloads*. In all panels, *LM Sentiment* and *Harvard Sentiment* sentiments are normalized to one in 2010 within each group, one year before the publication of Loughran and McDonald (2011). The dotted lines represent the 95% confidence limits.

Panel A: Top tercile machine downloads vs. bottom tercile machine downloads



Panel B: Top quartile machine downloads vs. the rest

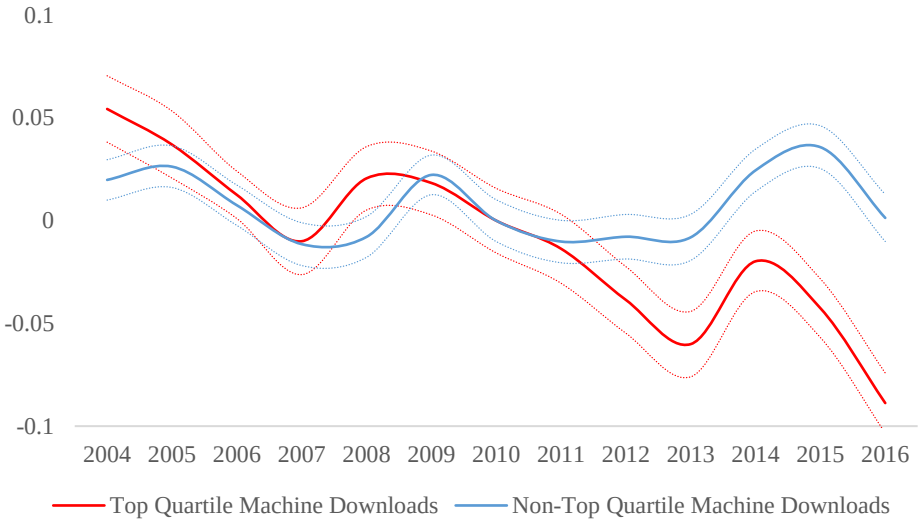


Table 1. Summary Statistics

This tables provide summary statistics. Filing level variables are based on the sample of SEC EDGAR 10-K and 10-Q filings from 2004 to 2016. Conference call level variables are based on the sample of the audio of corporate conference calls from 2010 to 2016. Firm-year level control variables are calculated annually using information available at the previous year-end. Variables are defined in Appendix A.

Variables	Mean	Median	Std	P25	P75	N
			Filing level			
Machine Downloads	4.729	4.508	1.763	3.296	6.377	324,607
Other Downloads	3.448	3.474	1.378	2.615	4.363	324,607
Total Downloads	5.09	4.915	1.609	3.829	6.535	324,607
%Machine Downloads	0.742	0.775	0.179	0.623	0.892	324,231
Machine Readability	-0.020	0.125	0.584	-0.224	0.359	199,421
LM – Harvard Sentiment	-2.413	-2.385	0.544	-2.747	-2.047	324,589
LM Sentiment	1.625	1.543	0.599	1.185	1.982	324,589
Harvard Sentiment	4.038	4.021	0.697	3.561	4.492	324,589
Litigious	0.965	0.82	0.537	0.593	1.177	324,589
Uncertainty	1.425	1.377	0.398	1.146	1.652	324,589
WeakModal	0.521	0.427	0.304	0.314	0.634	324,589
StrongModal	0.295	0.271	0.133	0.202	0.359	324,589
			Conference call level			
Emotion_Valence	0.331	0.375	0.261	0.227	0.498	43,462
Emotion_Arousal	0.647	0.650	0.138	0.557	0.740	43,462
			Firm-year level control variables			
Size	6.238	6.22	2.022	4.804	7.617	43,764
Tobin's Q	0.672	0.557	0.718	0.178	1.064	43,764
ROA	0.0491	0.101	0.271	0.028	0.163	43,764
Leverage	0.221	0.16	0.244	0.008	0.337	43,764
Growth	0.152	0.0736	0.42	-0.005	0.191	43,764
IndAdjRet	0.000	-0.001	0.039	-0.021	0.019	43,764
InstOwnership	0.482	0.528	0.359	0.080	0.816	43,764
Log(#analyst)	1.498	1.609	1.193	0	2.485	43,764
IdioVol	0.463	0.386	0.289	0.263	0.576	43,764
Turnover	2.150	1.619	1.960	0.826	2.791	43,764
Segment	5.323	5	3.564	2	7	43,764

Table 2. Determinants of Machine Downloads

This tables reports the determinants of *Machine Downloads* and *% Machine Downloads*. Variables are defined in Appendix A. *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	*Machine Downloads*			*% Machine Downloads*			*Machine Downloads*	
Size	0.135***	0.139***	0.040***	-0.028***	-0.029***	-0.008***	0.139***	0.040***
	(40.29)	(45.62)	(7.05)	(-26.71)	(-28.91)	(-6.01)	(45.62)	(7.05)
Tobin's Q	-0.048***	-0.066***	-0.022***	0.010***	0.011***	0.002	-0.066***	-0.022***
	(-9.41)	(-13.24)	(-3.38)	(6.51)	(8.09)	(1.25)	(-13.24)	(-3.38)
ROA	-0.011	-0.031***	-0.002	-0.013***	0.012***	0.013***	-0.031***	-0.002
	(-0.94)	(-2.68)	(-0.14)	(-3.78)	(3.15)	(3.46)	(-2.68)	(-0.14)
Leverage	0.085***	0.122***	0.055***	-0.019***	-0.020***	-0.010**	0.122***	0.055***
	(6.58)	(9.39)	(3.37)	(-5.25)	(-5.75)	(-2.48)	(9.39)	(3.37)
Growth	-0.078***	-0.068***	-0.024***	-0.004**	-0.008***	-0.011***	-0.068***	-0.024***
	(-13.69)	(-12.21)	(-3.63)	(-2.50)	(-5.31)	(-6.76)	(-12.21)	(-3.63)
IndAdjRet	-0.847***	-0.729***	-0.322***	0.217***	0.188***	0.084***	-0.729***	-0.322***
	(-15.75)	(-13.97)	(-6.00)	(15.44)	(14.22)	(7.31)	(-13.97)	(-6.00)
InstOwnership	-0.005	-0.024*	-0.026	0.038***	0.045***	0.028***	-0.024*	-0.026
	(-0.32)	(-1.66)	(-1.24)	(8.47)	(11.19)	(6.48)	(-1.66)	(-1.24)
Log(#analyst)	-0.008	-0.008	-0.021***	-0.004**	-0.005***	0.000	-0.008	-0.021***
	(-1.52)	(-1.54)	(-2.92)	(-2.46)	(-3.38)	(0.07)	(-1.54)	(-2.92)
IdioVol	0.091***	0.060***	-0.062***	-0.080***	-0.073***	-0.028***	0.060***	-0.062***
	(6.07)	(4.32)	(-4.37)	(-17.94)	(-18.71)	(-8.68)	(4.32)	(-4.37)
Turnover	0.022***	0.019***	0.022***	-0.006***	-0.005***	-0.006***	0.019***	0.022***
	(13.20)	(12.08)	(12.11)	(-12.11)	(-10.90)	(-14.55)	(12.08)	(12.11)
Segment	0.007***	0.007***	0.007***	-0.000	-0.001***	-0.001***	0.007***	0.007***
	(6.81)	(6.97)	(3.89)	(-0.55)	(-3.06)	(-2.94)	(6.97)	(3.89)
AIHedgeFund							0.728***	0.417**
							(4.52)	(2.54)
Observations	171,296	171,296	171,234	171,244	171,244	171,182	171,296	171,234

R-squared	0.924	0.926	0.941	0.658	0.690	0.808	0.926	0.941
Company FE	No	No	Yes	No	No	Yes	No	Yes
Industry FE	No	Yes	No	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 3. Machine Downloads and Machine Readability

This table examines the relation between the machine readability of a firm's filing and the machine downloads of the firm's past filings. Variables are defined in Appendix A. In Panel B, *Machine Downloads (Alt. def.)* and *Other Downloads (Alt. def.)* are alternative definitions of *Machine Downloads* and *Other Downloads* based on a criterion to classify machine visits in Loughran and McDonald (2017). Panel C reports the underlying components of *Machine Readability*, including *Table Extraction* (the ease of separating tables from text), *Number Extraction* (the ease of extracting numbers from text), *Table Format* (the ease of identifying the information contained in the table), *Self-Containedness* (whether a filing includes all needed information, i.e., without relying on external exhibits), and *Standard Characters* (the proportion of characters that are standard ASCII characters). Each attribute is standardized. In all panes, *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

Panel A: Machine readability

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable			*Machine Readability*			
Machine Downloads	0.076***	0.075***	0.060***	0.078***		
	(13.89)	(17.45)	(10.33)	(15.93)		
Other Downloads	0.005	0.002	-0.007	-0.006		
	(1.15)	(0.47)	(-1.44)	(-1.33)		
% Machine Downloads					0.121***	0.173***
					(3.91)	(6.39)
Total Downloads					0.053***	0.074***
					(10.27)	(16.26)
Size			0.004	0.021***	0.004	0.021***
			(1.05)	(2.66)	(0.90)	(2.64)
Tobin's Q			-0.006	-0.008	-0.006	-0.008
			(-0.92)	(-1.00)	(-0.91)	(-0.99)
ROA			0.056***	0.009	0.057***	0.010
			(3.15)	(0.49)	(3.19)	(0.52)
Leverage			-0.087***	-0.037*	-0.086***	-0.037*
			(-4.62)	(-1.67)	(-4.60)	(-1.67)
Growth			-0.017**	0.010	-0.017**	0.010
			(-2.34)	(1.27)	(-2.34)	(1.26)
IndAdjRet			0.033	0.013	0.038	0.015
			(0.52)	(0.20)	(0.60)	(0.24)
InstOwnership			0.050***	-0.038	0.051***	-0.039
			(2.69)	(-1.50)	(2.73)	(-1.54)
Log(#analyst)			0.005	0.000	0.005	0.000
			(0.79)	(0.02)	(0.81)	(0.06)
IdioVol			-0.072***	0.015	-0.074***	0.015
			(-3.81)	(0.86)	(-3.90)	(0.85)
Turnover			-0.002	-0.007***	-0.002	-0.007***
			(-1.17)	(-3.16)	(-1.12)	(-3.06)
Segment			0.004***	-0.003	0.004***	-0.003
			(3.05)	(-1.42)	(3.03)	(-1.43)

Observations	198,358	199,241	150,425	150,346	150,377	150,298
R-squared	0.082	0.363	0.084	0.357	0.084	0.357
Company FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Alternative specifications

	(1)	(2)	(3)	(4)
Dependent Variable	*PCA Machine Readability*		*Machine Readability*	
Machine Downloads	0.131***	0.162***		
	(11.18)	(16.14)		
Other Downloads	-0.047***	-0.046***		
	(-4.75)	(-5.88)		
Machine Downloads (Alt. def.)			0.052***	0.064***
			(9.51)	(13.72)
Other Downloads (Alt. def.)			-0.010	-0.000
			(-1.51)	(-0.05)
Size	-0.036***	0.019	0.005	0.021***
	(-4.02)	(1.34)	(1.20)	(2.65)
Tobin's Q	-0.013	-0.022	-0.007	-0.008
	(-0.90)	(-1.43)	(-0.97)	(-0.98)
ROA	0.245***	0.054	0.056***	0.010
	(6.15)	(1.52)	(3.15)	(0.54)
Leverage	-0.171***	-0.040	-0.085***	-0.038*
	(-4.60)	(-0.98)	(-4.55)	(-1.70)
Growth	-0.092***	-0.002	-0.017**	0.009
	(-5.80)	(-0.12)	(-2.34)	(1.21)
IndAdjRet	0.432***	0.144	0.031	0.015
	(3.66)	(1.28)	(0.48)	(0.24)
InstOwnership	0.108***	0.009	0.051***	-0.037
	(2.75)	(0.19)	(2.71)	(-1.44)
Log(#analyst)	-0.012	-0.005	0.005	0.000
	(-0.88)	(-0.35)	(0.77)	(0.01)
IdioVol	-0.360***	-0.044	-0.072***	0.014
	(-10.11)	(-1.53)	(-3.78)	(0.80)
Turnover	-0.018***	-0.015***	-0.002	-0.007***
	(-4.06)	(-3.47)	(-1.07)	(-3.25)
Segment	0.012***	-0.001	0.004***	-0.003
	(3.78)	(-0.21)	(3.06)	(-1.46)
Observations	139,436	139,330	150,425	150,346
R-squared	0.089	0.336	0.084	0.357
Company FE	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes

Panel C: Components of *Machine Readability*

	(1)	(2)	(3)	(4)	(5)
	Machine Readability				
Dependent Variable	*Table Extraction*	*Number Extraction*	*Table Format*	*Self-Containedness*	*Standard Characters*
Machine Downloads	0.051***	0.028***	0.026***	0.161***	0.125***
	(6.02)	(3.47)	(2.88)	(21.80)	(14.68)
Other Downloads	0.018**	-0.011	0.022**	-0.036***	-0.040***
	(2.37)	(-1.49)	(2.51)	(-6.69)	(-6.08)
Size	0.037***	0.043***	0.012	0.033***	-0.032**
	(2.67)	(3.50)	(0.85)	(3.44)	(-2.53)
Tobin's Q	-0.015	-0.054***	0.010	-0.006	0.028**
	(-1.00)	(-3.97)	(0.63)	(-0.52)	(2.26)
ROA	0.031	0.030	-0.006	-0.038	0.040
	(0.92)	(0.88)	(-0.15)	(-1.55)	(1.30)
Leverage	0.015	0.020	-0.060	-0.018	-0.117***
	(0.37)	(0.62)	(-1.36)	(-0.63)	(-3.29)
Growth	0.010	0.005	0.022	0.007	-0.007
	(0.71)	(0.38)	(1.51)	(0.58)	(-0.47)
IndAdjRet	-0.051	0.088	-0.075	-0.197***	0.253***
	(-0.48)	(0.85)	(-0.61)	(-2.63)	(2.81)
InstOwnership	-0.095**	-0.017	-0.063	-0.015	0.046
	(-2.05)	(-0.44)	(-1.24)	(-0.47)	(1.15)
Log(#analyst)	0.003	0.006	0.009	-0.009	-0.009
	(0.20)	(0.44)	(0.57)	(-0.96)	(-0.81)
IdioVol	0.005	-0.020	0.054	0.043**	-0.018
	(0.17)	(-0.70)	(1.51)	(2.12)	(-0.76)
Turnover	-0.008**	-0.003	-0.006	-0.007**	-0.012***
	(-2.07)	(-0.81)	(-1.36)	(-2.19)	(-3.26)
Segment	-0.002	0.006	-0.011***	0.004*	-0.013***
	(-0.67)	(1.55)	(-2.75)	(1.75)	(-3.98)
Observations	149,484	150,346	149,484	150,245	140,061
R-squared	0.471	0.389	0.439	0.306	0.344
Company FE	Yes	Yes	Yes	Yes	Yes
Industry FE	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes

Table 4. Consequences of Machine Reading: Time to the First Trade

This table examines the relation between the time to the first trade after a firm's filing is publicly released the machine downloads of the firm's past filings, and how the machine readability of the filings affects such a relation. All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	*Time to the First Trade*				*Time to the First Directional Trade*			
Machine Downloads	-8.353**	-4.857*	-7.347**	-3.398	-12.365***	-7.540***	-12.374***	-7.258**
	(-2.56)	(-1.68)	(-2.19)	(-1.14)	(-3.94)	(-2.71)	(-3.87)	(-2.55)
Machine Downloads × *Machine Readability*			-3.761**	-3.887***			-2.815*	-2.127*
			(-2.46)	(-2.84)			(-1.87)	(-1.67)
Machine Readability			-6.540	-5.980			-5.695	-8.709
			(-0.99)	(-0.92)			(-0.91)	(-1.46)
Other Downloads	15.342***	3.499	15.151***	1.304	13.961***	3.885*	13.436***	2.336
	(5.29)	(1.42)	(5.06)	(0.51)	(4.95)	(1.72)	(4.67)	(1.00)
Size	-50.806***	-38.789***	-51.227***	-38.997***	-48.121***	-35.627***	-48.908***	-35.923***
	(-23.29)	(-10.29)	(-22.35)	(-9.82)	(-21.67)	(-9.93)	(-21.06)	(-9.49)
Tobin's Q	-6.457*	-12.396***	-5.779	-12.621***	-4.747	-13.633***	-3.847	-13.359***
	(-1.76)	(-2.99)	(-1.54)	(-2.89)	(-1.34)	(-3.57)	(-1.07)	(-3.30)
ROA	-34.069***	-4.892	-30.756***	-4.168	-34.933***	-6.956	-33.623***	-5.071
	(-4.13)	(-0.50)	(-3.61)	(-0.40)	(-4.50)	(-0.86)	(-4.23)	(-0.59)
Leverage	12.422	8.196	7.754	-0.451	6.006	4.097	3.909	-0.921
	(1.30)	(0.75)	(0.77)	(-0.04)	(0.66)	(0.41)	(0.42)	(-0.09)
Growth	16.116***	-1.510	15.103***	-0.341	17.820***	-1.199	17.403***	0.218
	(4.53)	(-0.36)	(3.99)	(-0.08)	(5.52)	(-0.31)	(5.09)	(0.05)
IndAdjRet	2.186	-7.888	-8.375	0.315	0.160	-13.379	-16.519	-17.567
	(0.06)	(-0.23)	(-0.23)	(0.01)	(0.00)	(-0.42)	(-0.49)	(-0.52)
InstOwnership	-39.142***	14.042	-41.458***	10.546	-33.161***	5.286	-34.708***	4.926
	(-3.62)	(1.07)	(-3.72)	(0.76)	(-3.09)	(0.41)	(-3.16)	(0.37)
Log(#analyst)	-6.209*	-8.422**	-5.999	-8.360**	-5.698	-4.882	-5.421	-4.682
	(-1.74)	(-2.18)	(-1.63)	(-2.07)	(-1.61)	(-1.31)	(-1.49)	(-1.22)
IdioVol	15.150*	-8.231	12.112	-11.668	0.438	-19.451**	-1.904	-19.783**
	(1.73)	(-0.96)	(1.34)	(-1.29)	(0.05)	(-2.46)	(-0.23)	(-2.40)

Turnover	-14.489***	-7.802***	-14.536***	-7.706***	-11.946***	-6.787***	-11.854***	-6.668***
	(-12.25)	(-6.55)	(-11.77)	(-6.19)	(-9.65)	(-5.91)	(-9.25)	(-5.51)
Segment	-0.588	0.984	-0.122	0.476	-0.945	1.220	-0.484	0.278
	(-0.76)	(1.07)	(-0.15)	(0.48)	(-1.23)	(1.36)	(-0.61)	(0.29)
Observations	161,749	161,664	144,281	144,193	161,749	161,664	144,281	144,193
R-squared	0.116	0.269	0.118	0.272	0.120	0.285	0.122	0.286
Company FE	No	Yes	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 5. Machine Downloads and Sentiment: Loughran and McDonald (2011) Publication

This table reports the impact of the publication of Loughran and McDonald (2011) on the relation between the sentiment of a firm's filing and the machine downloads of the firm's past filings. Control variables include *Other Downloads*, *Size*, *Tobin's Q*, *ROA*, *Leverage*, *Growth*, *IndAdjRet*, *InstOwnership*, *Log(#analyst)*, *IdioVol*, *Turnover*, and *Segment.* All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	*LM – Harvard Sentiment*		*LM Sentiment*		*Harvard Sentiment*	
Machine Downloads × Post	-0.072***	-0.079***	-0.062***	-0.050***	0.010	0.029***
	(-6.95)	(-8.94)	(-4.98)	(-4.99)	(0.76)	(2.65)
Machine Downloads	-0.007	-0.011**	-0.009	-0.019***	-0.002	-0.008
	(-1.17)	(-2.46)	(-1.18)	(-3.72)	(-0.23)	(-1.43)
Observations	158,578	158,515	158,578	158,515	158,578	158,515
R-squared	0.217	0.568	0.241	0.632	0.208	0.590
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes
Company FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Table 6. Machine Downloads and Other Tones: Loughran and McDonald (2011) Publication

This table reports the impact of the publication of Loughran and McDonald (2011) on the relation between the various tones of a firm's filing and the machine downloads of the firm's past filings. Control variables include *Other Downloads, Size, Tobin's Q, ROA, Leverage, Growth, IndAdjRet, InstOwnership, Log(#analyst), IdioVol, Turnover,* and *Segment.* All variables are defined in Appendix A. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	*Litigious*		*Uncertainty*		*Weak Modal*		*Strong Modal*	
Machine Downloads × Post	-0.056***	-0.057***	-0.016**	-0.021***	-0.028***	-0.034***	-0.008***	-0.007***
	(-5.38)	(-6.02)	(-2.01)	(-3.49)	(-4.85)	(-8.86)	(-4.39)	(-4.39)
Machine Downloads	0.011*	0.007	-0.006	-0.009***	-0.018***	-0.021***	-0.003**	-0.004***
	(1.71)	(1.44)	(-1.33)	(-3.05)	(-5.39)	(-10.05)	(-2.19)	(-4.98)
Observations	158,578	158,515	158,578	158,515	158,578	158,515	158,578	158,515
R-squared	0.188	0.509	0.196	0.600	0.238	0.624	0.277	0.571
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Company FE	No	Yes	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No	Yes	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 7. Machine Downloads and Managers' Emotion during Conference Calls

This table examines the relation between the manager's speech emotion during conference calls and the machine downloads of the firm's past filings. Control variables include *Other Downloads, Size, Tobin's Q, ROA, Leverage, Growth, IndAdjRet, InstOwnership, Log(#analyst), IdioVol, Turnover,* and *Segment* as in the previous tables. Columns (4) and (8) further include *EarningsSurprise* as an additional control. All variables are defined in Appendix A. The sample consists of audio of conference calls between January 2010 and December 2016. The *t*-statistics, in parentheses, are based on standard errors clustered by firm. ***, **, * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two-tailed), respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	*Emotion-Valence*				*Emotion-Arousal*			
Machine Downloads	0.043***	0.035***	0.042***	0.042***	0.004*	0.003	0.005**	0.007**
	(11.40)	(8.13)	(11.14)	(8.84)	(1.79)	(0.94)	(2.28)	(2.49)
Other Downloads	-0.017***	-0.014***	-0.017***	-0.012***	-0.006***	0.000	-0.006***	-0.006***
	(-5.74)	(-4.32)	(-5.67)	(-3.12)	(-3.65)	(0.19)	(-3.71)	(-2.92)
Observations	43,336	41,340	41,224	27,437	43,336	41,340	41,224	27,437
R-squared	0.389	0.189	0.383	0.388	0.395	0.132	0.395	0.469
Control Variables	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Company FE	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Industry FE	No	Yes	No	No	No	Yes	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes