How to Talk When Machines are Listening: Corporate Disclosure in the Age of AI

Sean Cao
J. Mack Robinson College of Business
Georgia State University
scao@gsu.edu

Wei Jiang
Columbia Business School and NBER
wj2006@columbia.edu

Baozhong Yang
J. Mack Robinson College of Business
Georgia State University
bzyang@gsu.edu

Alan L. Zhang
J. Mack Robinson College of Business
Georgia State University
lzhang27@gsu.edu

This draft: May 2020

Preliminary and for Conference Review Only

Please Do Not Distribute

How to Talk When Machines are Listening: Corporate Disclosure in the Age of AI

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 AI readership, proxies by machine downloads, motivates firms to prepare filings that are more friendly to machine parsing, scripting, and processing. Moreover, firms with high expected machine downloads manage textual sentiment in ways catered to AI readers by differentially avoiding words that are perceived as negative by algorithms, as compared to those by human readers. The publication of Loughran and McDonald (2011) is instrumental in attributing the change in the measured sentiment to AI readership. While existent research has explored 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.

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 of Berkshire Hathaway 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's 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 potential readership. 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] 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 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 tools, for example, automated processes that identify positive, negative and neutral opinions in a body of text. While the literature is catching up with and guiding investors' rising aptitude to apply machine-learning and computational tools to extract 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 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

[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.)

variable *Machine Downloads* is constructed by tracking IP addresses that conduct downloads in batches. In regressions the variable uses historical information of downloads of the firms' filings to capture the propensity of machine downloads of a given filing, to avoid a look-head bias in a variable that is otherwise mostly out of the control of the firm. 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 seven elements, identified by the recent and burgeoning literature (see Section 2), as affecting the ease for machine parsing, scripting, and synthesizing. The default measure for *Machine Readability* takes the first principal component of the seven underlying elements. We show that, in the cross-section of filings with firm and year fixed effects, a change from the bottom to the top quintile of expected machine downloads is associated with 0.64 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.

After establishing a positive association between a high AI reader base and more machine-friendly disclosure documents produced by the firm, we further explore how firms manage "sentiment" perceived by machines. It is well-documented that corporate disclosures attempt to strike the right tones with (human) readers by conveying positively-biased sentiments 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") 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 and negative words, but also the word list that came with the paper has served as a leading lexicon for algorithms to sort out sentiments.[2] The differences in both the timeline and the context of the two dictionaries allow us to identify the impact of AI readership on sentiment management by corporations.

As a first step, we establish that firms which expect high (i.e., the top quintile) machine downloads and the other (control) firms do not exhibit differences in their sentiments as measured by either the Harvard Dictionary (which has served human readers for a long time) or the LM Dictionary up to 2010 (the year before publication). Such a parallel pre-trend between firms with high and low machine downloads suggests that prior to the LM Dictionary, firms did not tune the sentiment of their reports differentially in measured ways for machine readers. Post-2011 saw a structural break: Firms with high AI readership significantly reduced their uses of negative words from the LM Dictionary, while there is little change for the control group. Moreover, the two groups remain similar in their sentiments as measured by the Harvard Dictionary in the post period. If we construct an *LM – Harvard Sentiment* measure to proxy the difference in perception by AI vs. human readers, we also observe a parallel pre-trend between firms with high and low AI readership prior to the publication of Loughran and McDonald (2011), and a significant divergence afterward. Given the quasi-randomness of the exact timing of publication, the change in the LM sentiment measure is more likely to be attributed to firms' catering to its AI readers than alternative forces.

[2] For example, 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.

Finally, we demonstrate the complementarity between machine readability and sentiment given that algorithm-driven sentiment should be conveyed more clearly if a report is more machine-friendly to start with. Indeed, we find that companies whose filings are more machine readable are more likely to adjust their sentiment writing after the publication of the LM Dictionary, as they anticipate that files downloaded by machines are more likely to be parsed and analyzed against the list of words provided in the new dictionary.

II. Data and Variables

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 1984, 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, the Central Index Key (CIK) used in the SEC's computer systems to identify filers, and accession numbers that uniquely match a particular company's specific SEC filing. 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.[3] This procedure yields a total of 410,931 filings (111,683 10-K and 299,248 10-Q). After matching to CRSP/Compustat, our final sample of raw filings consists of 199,244 filings (49,993 10-K and 149,251 10-Q), filed by 6,710 unique CIKs, between 2003 and 2015.[4]

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 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

Several constructed variables are instrumental in our analyses. 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. 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.[5] Some recent studies

[3] We do not include amendments and other variant filings because these documents likely mirror the original filings.
[4] We plan to extend the sample to more recent years. The CRSP/Compustat match rate of 48% is comparable to that in the literature, such as 46% in Loughran and McDonald (2014).
[5] "Cohen's Point72 Hires 30 People for Big Data Investing," Simone Foxman, Bloomberg, Mach 10, 2015.

also provide evidence that investment companies rely on machine downloads of EDGAR filings for some of their trading strategies. Crane, Crotty, and Umar (2019) find that hedge funds that employ robotic downloads perform better than those that do not. Cao, 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 daily as a machine (i.e., robot) visitor, the same criterion as used by Lee, Ma, and Wang (2015).[6] In addition, we include requests that are attributed to web crawlers in the SEC Log File Data as machine-initiated. All other 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.[7]

Figure 1 shows the exponential growth of machine downloads since 2004. The number of machine downloads of corporate 10-K and 10-Q filings increased from 413,573 in 2004 to 21,163,356 in 2015, a 50-fold increase. 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 45% to 91%.

[Insert Figure 1 here.]

To measure the propensity of machine downloads of a particular filing using ex ante information only, we construct the variable *# Prior Machine Downloads* as follows. For a firm's

[6] We also use an alternative approach to classify those daily IP addresses having more than 50 requests as robot visitors (a criterion used by Loughran and McDonald 2017). The results, appeared in robustness checks, are qualitatively similar.

[7] We use seven days because the majority of requests happened within the first week. We use alternative cutoffs including 14 days and 30 days as robustness checks.

(indexed by *i*) filing (indexed by *j*) on day *t*, *# Prior Machine Downloads* is the average number of machine downloads of the firm's historical filings that were filed during days [*t* – 390, *t* – 30] days (while we only include the machine downloads of a historical filing within seven days of posting on EDGAR, as we explained earlier). The variable is highly nonstationary (as shown in Figure 1) and skewed in the cross section. Therefore, we transform the variable in annual percentiles scaled to [0,1]. We label the scaled rank variable as *Prior Machine Downloads* and use it in our regression analyses.

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 (Hodge, Kennedy, and Maines, 2004, Allee, DeAngelis and Moon, 2018, Blankespoor, 2019, Blankespoor, deHaan, and Marinovic, 2020, Gao and Huang, 2020) has studied various concepts and proposed metrics of information processing costs, which could be related to either machine or human processing costs (or both). After reviewing the existing research, we create a machine readability measure by considering the following attributes that are most distinctly related to machine readability: (i) the ease of separating tables from text; (ii) the ease of extracting numbers from text; (iii) the ease of identifying the information contained in the table; (iii) the completeness and consistency of tabular data; (iv) whether a filing includes all needed information(i.e., without relying on external exhibits); (v) whether a filing includes programmatically unfriendly parts such as a PDF or image; (vi) the proportion of tokens that are a valid word or number; and (vii) the proportion of words that are correctly spelled.[8] In our main

[8] We thank Robbie Moon for sharing part of the data used in Allee, DeAngelis and Moon (2018).

regression, we use the first principal component of the seven attributes as the summary measure of *Machine Readability*, but present sensitivity checks using the underlying components.

Figure 2 provides some intuition behind the *Machine Readability* variable by showing two sample filings: one with a low score (-4.16, or 2.8 standard deviations below the mean) by CARNIVAL PLC in 2007, and one with a high readability score (2.00, or 1.36 standard deviations above the mean) by VIASAT INC in 2012. We contrast the excerpts of the two filings as follows.

[Insert Figure 2 here.]

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 identify such a "table." Next, the filing refers to more than ten external exhibits (e.g., "d0770672_ex32-6.txt"), which are not included in the filing. Furthermore, the filing contains PDF files that account for over 20% of the length of the raw filing; and these PDF files are binary codes that are difficult to be processed programmatically. The excerpt of the second filing, in contrast, contains tags such as <Table>, <TR>, and <TD>, 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 contain a PDF file or an image, nor does it refer to an external exhibit.

The third class of key variables concerns measures for "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 perceptions 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 we use is the Harvard 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 (i.e., the total number of words in the document). The resulting measure becomes the *Harvard Sentiment*.

The second lexicon is developed by Loughran and McDonald (2011, LM hereafter), 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. We term the resulting measure the *LM Sentiment*.[9] Finally, we form the difference, *LM – Harvard Sentiment,* to capture the contrast.

As usual, the firm characteristics variables (as control variables) are retrieved or based on information from standard databases, such as CRSP/Compustat, and Thomson Reuters Ownership Database. *Size* is the market capitalization in logarithm. *BM* is the ratio of book to market value of firm equity. *ROA* is return on assets, or the ratio of net income to assets. *Momentum* is the cumulative stock return from month $m - 12$ to $m - 1$ where m is the filing month. *# Institutional Owner* is the number of institutional owners at the most recent quarter-end based on last quarter-end Form 13F filings. *Product Market Fluidity* is a proxy for the degree of competitive threat and product market disruption surrounding a firm, following Hoberg, Phillips and Prabhala (2014). It captures changes in rival firms' products relative to the firm's products and derives from firms' product text descriptions in 10-K filings. *Investment-to-Assets* is the ratio of the changes in property, plant and equipment, and inventories to assets. *Book Leverage* is the ratio of assets to book equity. *Advertising* and *R&D* are the name-sake expense items from Compustat, and are

[9] We consider only the negative sentiment following LM, who find that positive sentiment is not as informative. Replacing the negative sentiment measure by a net sentiment measure does not change our results qualitatively.

assigned to zero if missing. *Kaplan-Zingales Index* is the financial constraints measuring the reliance on external financing, following Kaplan and Zingales (1997). All control variables are constructed annually except *Momentum*, *ROA*, and *# Institutional Owner*. All potentially unbounded variables are winsorized at the 1% extremes.

The summary statistics are reported in Table 1. 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 149,771 filings (36,551 10-K and 113,220 10-Q).

[Insert Table 1 here.]

III. Machine Readership and Corporate Disclosure Decisions: Empirical Analyses

A. 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 hypothesis that companies adjust the way they talk 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 2 reports the results from the following regression at the filing level, indexed by firm(i)-filing(j)-date(t), with both firm and year fixed effects:

$$Machine\ Readability_{i,j,t} = \beta_1 Prior\ Machine\ Downloads_{i,j,t} + \beta_2 Prior\ Other\ Downloads_{i,j,t} + \boldsymbol{\gamma}\ Control_{i,year} + \alpha_i + \alpha_{year} + \varepsilon_{i,j,t}. \quad (1)$$

[Insert Table 2 here.]

Table 2 shows that higher propensity of machine downloads for a filing of a company, measured by *Prior Machine Downloads,* is indeed associated with the report being more machine-

reading friendly. Moreover, the coefficients on *Prior Machine Downloads* are significant at the 1% level, and are almost invariant with or without the control variables, indicating that machine downloads have little confounding effect with other firm characteristics. Because both variables do not have a natural unit, we interpret the relation in terms of their standard deviations: A change from the bottom quintile to top quintile in *Prior Machine Downloads* is associated with 0.64 standard-deviation change in *Machine Readability.*

It is important to note that the hypothesis that firms accommodate machine readers should predict no positive relation between *Machine Readability* and *Prior Other Download*. Hence the other downloads serve as a natural placebo test. Such a no-result is also confirmed by data. Coefficients on *Prior Other Downloads* turn out to be insignificant and negative.

B. Managing sentiment with machine readers

While truthfulness in disclosure reports is expected and required, it is well known that corporate disclosures are positively biased usually 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 is meant for human readers and for general purposes, and sets up 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, 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.[10]

The timeline of Harvard General Inquirer dictionary (existed since 1996) and the Loughran-McDonald dictionary (since 2011)[11], and their differential adoption by human vs. machine readers, provide a unique setting for us to test how the writing of corporate filings adjusts to AI readers. In particular, we consider the following regression at the filing level, indexed by firm(i)-filing(j)-date(t), with both firm and year fixed effects:

$$Negative\ Sentiment_{i,j,t} = \beta_1 Prior\ Machine\ Downloads_{i,j,t} + \beta_2 Prior\ Other\ Downloads_{i,j,t} + \boldsymbol{\gamma}\ Control_{i,year} + \alpha_i + \alpha_{year} + \varepsilon_{i,t}. \quad (2)$$

In the equation above, we consider three measures for the dependent variable *Negative Sentiment.* The first is the *Harvard Sentiment*, constructed as the number of Harvard General

[10] 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.
[11] The paper was in public distribution, e.g., posted on the SSRN, since 2009.

Inquirer negative words in a filing divided by the total number of words in the filing. The second is the *LM Sentiment,* defined as the number of Loughran-McDonald (LM) Financial-Negative words in a filing divided by the total number of words in the filing. Finally, we take the difference of the two to form *LM – Harvard* Sentiment. Results are reported in Table 3.

[Insert Table 3.]

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.* This is precisely what Table 3 shows. The coefficients associated with *Prior Machine Downloads* are negative and significant (at the 1% level) in predicting *LM Sentiment*, but the effect is indistinguishable from zero, in terms of both magnitude and significance, in predicting *Harvard Sentiment.* Not surprisingly, the relation between *LM – Harvard Sentiment* and *Prior Machine Downloads* is qualitatively similar to that of *LM Sentiment.*

In contrast, there is no negative relation between the negative sentiments and the amount of *Prior Other Downloads.* In fact, the coefficients on *Prior Other Downloads* are all positive and significant.

The relation revealed in Table 3 supports the hypothesis that corporate writing has adjusted to the availability of the Loughran-McDonald dictionary. At the same time, however, a reverse relation is almost equally plausible that the dictionary summarizes the practice that was already in place, and serves as a concurrent side-show. Under the first hypothesis, the relation between *LM Sentiment* and machine downloads should be present only after the publication of Loughran and McDonald (2011). Such an exclusive effect is confirmed by results in Table 4, where we added an interaction term *Prior Machine Downloads* × *Post* to equation (2), where *Post* is an indicator

variable equal to one for filings in 2012 and onwards, and zero otherwise (filings in 2011 are excluded from the analysis). Because year fixed effects are incorporated, the variable *Post* is subsumed.

[Insert Table 4]

Table 4 shows an unambiguous contrast before and after 2011, the year when Loughran and McDonald (2011) was published. Post 2011, a change from the bottom quintile to the top quintile in *Prior Machine Downloads* is associated with a 0.16% decrease in LM sentiment, which is about 10% of the sample mean of *LM Sentiment,* or 0.25 standard deviations. The relation is significant at the 1% level. The relation between *Harvard Sentiment* and *Prior Machine Downloads* is insignificant when control variables are incorporated. In the early period, however, the relation between *LM Sentiment* and *Prior Machine Downloads* was actually significantly positive, and there is no statistical difference in the effects regarding *Harvard Sentiment* and *LM Sentiment.*

It remains possible that results in Table 4 merely reflects a general trend of a strengthening in the relation between the machine downloads and avoiding using words that are perceived to have negative annotations in the finance context. We thus resort to Figure 3 to illustrate the structural break, instead of a pre- 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 quintile of *Prior Machine Downloads* and the rest four quintiles. We then plot in Figure 3 the time series of both groups of filings' tendency to use negative words that are distinct in the finance context, in excess of the general-context negative words.

[Insert Figure 3 here.]

Figure 3, Panel A shows parallel pre-trends till 2011 and then a clear divergence afterward. Before 2010, the *LM – Harvard* sentiment have the same trends for the group of top quintile in *Prior Machine Downloads* and the group of the other four quintiles. Beginning in early 2010, one year after the working paper version of Loughran and McDonald (2011) began to circulate (and became public on the SSRN), the sentiment trends for two groups started to diverge and the divergence became more apparent after 2011 when the paper was eventually published. The gap between *LM* and *Harvard* sentiment is also mostly driven by a dramatic decline of the *LM* negative sentiment since 2010, which declined by around 10% from 2010 to 2014. Given the quasi-randomness of the exact year of publication, it is unlikely that the publication of Loughran and McDonald (2011) had the perfect timing to ride on the structural break.

Table 5 demonstrates the difference-in-difference in a regression framework. Since Loughran and McDonald (2011) was first posted on SSRN in February 2009, we create several dummy variables to capture the trends up to and after the SSRN and journal publications. Specifically, *Between SSRN and Publish* in an indicator variable equal to one for filings in 2009 and 2010. *PublishYear* is an indicator variable equal to one for filings in 2011 alone. *PostPublish_1Year* is an indicator variable equal to one for filings in 2012 alone. *PostPublish_2Year+* is an indicator variable equal to one for filings in 2013 and onwards. Finally, *PreSSRN_1Year* is an indicator variable equal to one for filings in 2008 alone. The results in Table 5 show that the correlation between *LM – Harvard Sentiment* with *Prior Machine Downloads* turned negative after the paper posted to SSRN and the negative relation strengthened after the paper's journal publication. Importantly, an insignificant coefficient of *Prior Machine Downloads* × *PreSSRN_1Year* indicates that *LM – Harvard Sentiment* did not vary with machine downloads, in other words, there was no pre-tend in the relation of the two variables.

[Insert Table 5 here.]

C. Complementarity between machine readability and sentiment

Algorithm-measured sentiment could be conveyed more clearly if a report is more scriptable to start with. Thus we further hypothesize a complementarity between *Machine Readability* and *LM Sentiment*. That is, the positive relation between *LM Sentiment* and *Prior Machine Downloads* should be stronger in the subsamples of firms whose reports are more machine readable. In Table 6, we repeat the regression in Table 4 except for separately samples with high and low machine readability. Each year, a firm's filing on day zero is sorted into two groups by the average *Machine Readability* of the firm's historical 10-K and 10-Q filings that were filed during days [-390, -30]. The group above (below) the median breakpoint is labeled as "High" ("Low"). The hypothesis is that companies with high machine readability are more likely to adjust their writing after the publication of LM dictionaries, with the expectation that files downloaded by machines are more likely to be parsed and analyzed against the list of words provided in the new dictionary.

[Insert Table 6]

Table 6 confirmed the expected cross-sectional heterogeneity. Post 2011, the relation between *Prior Machine Downloads* and *LM – Harvard Sentiment* is about 40% stronger in the high *Machine Readability* subsample than that in the lower subsample, and the difference is significant at the 10% level.

IV. Conclusion

This paper presents the first study showing how corporate disclosure writing 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. Moreover, firms manage sentiment 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.

References

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

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.

Bodnaruk, Andriy, Tim Loughran, and Bill McDonald, 2015, Using 10-K text to gauge financial constraints, *Journal of Financial and Quantitative Analysis* 50, 623–646.

Cao, Sean, Kai Du, Baozhong Yang, and Alan 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.

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.

Crane, Alan D. and Kevin Crotty and Tarik Umar, 2019, Hedge funds and public information acquisition, Working paper, Rice University.

Frésard, Laurent, Gerard Hoberg, and Gordon M Phillips, 2019, Innovation Activities and Integration through Vertical Acquisitions, *Review of Financial Studies* forthcoming.

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.

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.

Hirschberg, Julia, and Christopher D. Manning, 2015, Advances in natural language processing, *Science* 349, 261–266.

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies and competition in mergers and acquisitions: A text-based analysis, *Review of Financial Studies* 23, 3773–3811.

Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of Political Economy* 124, 1423–1465.

Hoberg, Gerard, and Gordon Phillips, 2018, Conglomerate industry choice and product language, *Management Science* 64, 3735–3755.

Hoberg, Gerard, and Gordon Phillips, 2019, Text-based industry momentum, *Journal of Financial and Quantitative Analysis* forthcoming.

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala, 2014, Product market threats, payouts, and financial flexibility, *Journal of Finance* 69, 293–324.

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.

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

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

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

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

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

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 2006 to 2015. 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.

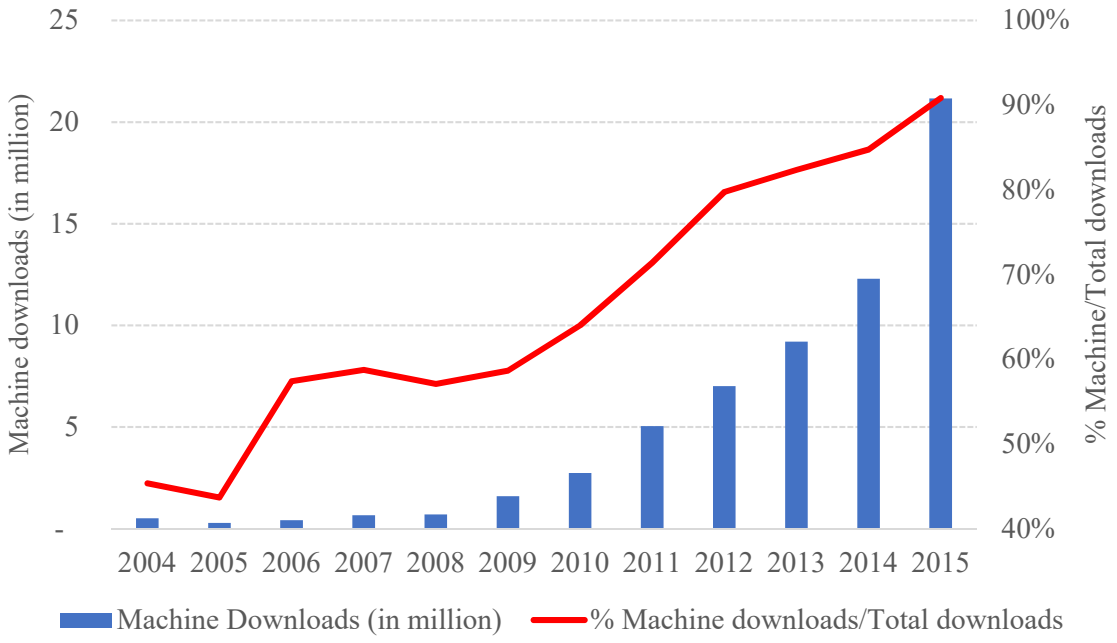


Figure 2. Excerpts of Two 10-K Filings

This figure shows two sample fillings, one with a low *Machine Readability* score (-4.16, or 2.8 standard deviation below the mean) by CARNIVAL PLC in 2007 and one with a high *Machine Readability* score (-2.00, or 1.36 standard deviation above the mean) by VIASAT INC in 2012. *Machine Readability* is the first principal component of a number of filing attributes, including the ease of separating tables from text, the ease of extracting numbers from text, the ease of identifying the information contained in tables, the completeness and consistency of tabular data, whether the filing includes all needed information without using external exhibits, whether the filing excludes programmatically unfriendly files such as PDF files or images in a disclosure, the proportion of tokens that are a valid word or number, and the proportion of words that are correctly spelled.

Excerpt 1. CARNIVAL PLC, CIK: 0001125259, February 12, 2007

Our revenues from the sale of passenger tickets are seasonal. Historically, demand for cruises has been the greatest during our third quarter, which includes the Northern Hemisphere summer months. This higher demand during the third quarter results in higher net revenue yields and, accordingly, the largest share of our net income is earned during this period. The seasonality of our results is increased due to ships being taken out of service for maintenance, which we typically schedule during non-peak demand periods. Substantially all of Holland America Tours' and Princess Tours' revenues and net income are generated from May through September in conjunction with the Alaska cruise season.

F-44

<PAGE>

Quarterly financial results for fiscal 2006 were as follows:

	Quarters Ended			
	February 28	May 31	August 31	November 30
	(in millions, except per share data)			
Revenues	$2,463	$2,662	$3,905	$ 2,809
Operating income	$ 349	$ 448	$1,340	$ 476
Net income(a)	$ 251	$ 380	$1,232	$ 416
Earnings per share				
Basic	$ 0.31	$ 0.47	$ 1.55	$ 0.53
Diluted	$ 0.31	$ 0.46	$ 1.49	$ 0.51
Dividends declared per share	$ 0.25	$ 0.25	$ 0.25	$ 0.275

(omitted)

<FILENAME>d0770672_ex32-6.txt
<DESCRIPTION>CERTIFICATIONS
<TEXT>

Exhibit 32.6

In connection with the Annual Report on Form 10-K for the year ended November 30, 2006 as filed by Carnival plc with the Securities and Exchange Commission on the date hereof (the "Report"), I certify pursuant to 18 U.S.C. Section 1350, as adopted pursuant to Section 906 of the Sarbanes-Oxley Act of 2002, that:

(1) The Report fully complies with the requirements of Section 13(a) or 15(d) of the Securities Exchange Act of 1934; and

(2) The information contained in the Report fairly presents, in all material respects, the financial condition and results of operations of Carnival plc.

Date: February 12, 2007

By: /s/ Gerald R. Cahill

Gerald R. Cahill
Executive Vice President
and Chief Financial and
Accounting Officer
</TEXT>
</DOCUMENT>
<DOCUMENT>
<TYPE>10-K
<SEQUENCE>21
<FILENAME>d70672_10-k.pdf
<DESCRIPTION>FORM 10-K
<TEXT>
<PDF>
begin 644 d70672_10-k.pdf
M)5!$1BTQ+C4-)>+CS],-"C4Y-2`P(&]B:CP\+TA;-C4V(#<Y-UTO3&EN96%R
M:7IE9"`Q+T4@-SDU,2],(#,R.#<T.2].(#$P-R]/(#4Y."]4(#,Q-C@P,3X^
M#65N9&]B:@T@("`@("`@("`@("`@("`@##0IX<F5F#0HU.34@,3@-"C`P,#`P
M,#`P,38@,#`P,#`@;@T*,#`P,#`P,30U,R`P,#`P,"!N#0HP,#`P,#`P-C4V
M(#`P,#`P(&X-"C`P,#`P,#$U-3@@,#`P,#`@;@T*,#`P,#`P,3<V-"`P,#`P

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

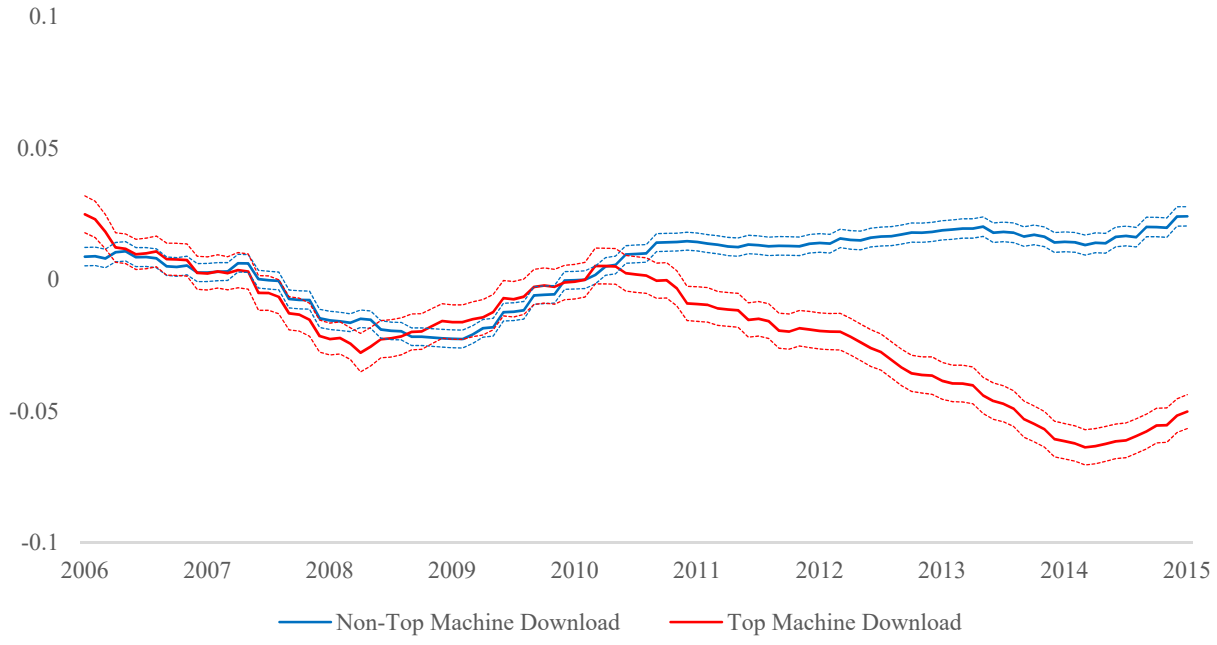
We amortize our acquired intangible assets from prior acquisitions over their estimated useful lives ranging from eight months to ten years. The decrease in amortization of acquired intangible assets of approximately $0.7 million in fiscal year 2012 compared to last fiscal year was a result of an approximately $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

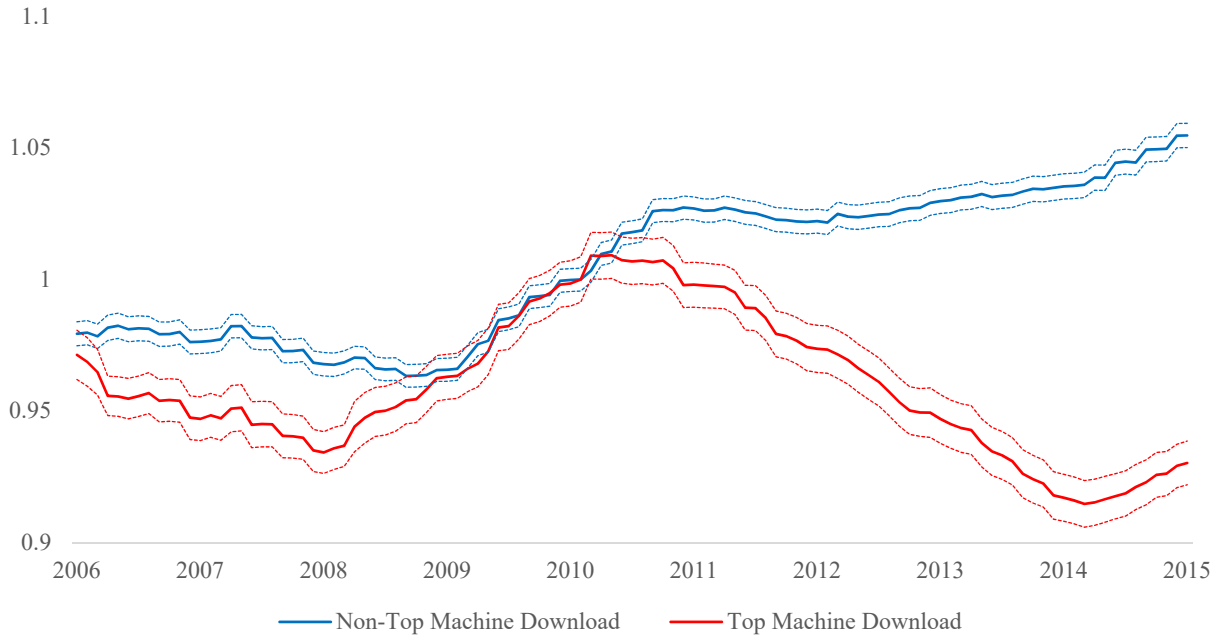
Figure 3. Sentiment Trend and Machine Downloads

This figure plots the trend of sentiment of 10-K and 10-Q filings at the monthly frequency from January 2006 to December 2015, and compares sentiment of firms with high machine downloads with that of the low group. There are three sentiment measures. *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. *LM Sentiment* is defined as the number of Loughran-McDonald (LM) Financial-Negative words in a filing divided by the total number of words in the filing. *LM – Harvard sentiment* is the difference of the two. At each month, a sentiment measure is calculated as the two-year moving average ending in the current month. Each chart presents the trend of two groups based on the quintiles of machine downloads. Each filing is sorted into either the "Top 20% Machine Download" group or the "Other" group based on machine downloads for the firms' historical filings filed during [-390, -30] days relative to the current filing. The quintiles are rebalanced annually. Panel A plots the *LM – Harvard sentiment*, and Panels B and C plot the two component sentiments. In all panels, *LM Sentiment* and *Harvard Sentiment* sentiments are normalized to one in January 2010, one year before the publication of Loughran and McDonald (2011). The time series start in 2006 due to the requirement of lagged two-year moving average. The dotted lines represent the 95% confidence intervals.

Panel A: LM – Harvard Sentiment



Panel B: LM Sentiment



Panel C: Harvard Sentiment

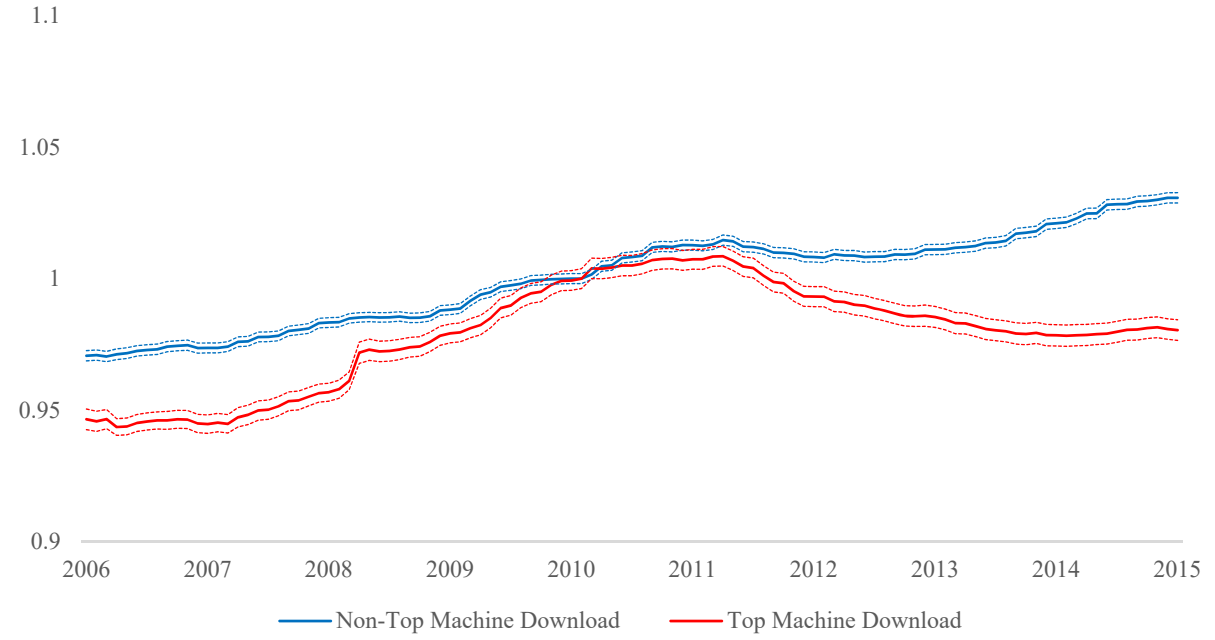


Table 1. Summary Statistics

This tables provide summary statistics on the sample of SEC EDGAR 10-K and 10-Q filings between January 2004 and December 2015. For a firm's filing on day zero, *# Prior Machine Downloads* (*# Prior Other Downloads*) is the average number of machine (other) downloads of the firm's historical 10-K and 10-Q filings (i.e., filings that were filed during days [-390, -30]). Each year, *# Prior Machine (Other) Downloads* is sorted into quintile ranks across filings by all firms and the ranks are scaled to [0, 1]; *Prior Machine Downloads* (*Prior Other Downloads*) is defined as the scaled ranks. *LM Sentiment* is defined as the ratio of the number of Loughran-McDonald (LM) Financial-Negative words in a filing to the total number of words in the filing, multiplied by 100. *Harvard Sentiment* is defined as the ratio of the number of Harvard General Inquirer Negative words in a filing to the total number of words in the filing, multiplied by 100. *LM – Harvard* is defined as *LM Sentiment* minus *Harvard Sentiment*. *Machine Readability* is the first principal component of a number of filing attributes, including the ease of separating tables from text, the ease of extracting numbers from text, the ease of identifying the information contained in tables, the completeness and consistency of tabular data, whether the filing includes all needed information without using external exhibits, whether the filing excludes programmatically unfriendly files such as PDF files or images in a disclosure, the proportion of tokens that are a valid word or number, and the proportion of words that are correctly spelled. *BM* is the ratio of book to market value of firm equity. *Size* is market capitalization in logarithm. *Momentum* is the cumulative stock return from month m – 12 to m – 1 where m is the filing month. *ROA* is the ratio of net income before extraordinary items to assets. *# Institutional Owner* is the number of institutional owners at the most recent quarter end based on Form 13F filings. *Product Market Fluidity* is a proxy for the degree of competitive threat and product market disruption surrounding a firm, following Hoberg, Phillips and Prabhala (2014). *Investment-to-Assets* is the ratio of the changes in property, plant and equipment, and inventories to assets. *Book Leverage* is the ratio of assets to book equity. *Kaplan-Zingales Index* is the financial constraints measuring the reliance on external financing, following Kaplan and Zingales (1997). *Advertising* is the advertising expense, and is assigned to zero if missing. *R&D* is the research and development expenses, and is assigned to zero if missing.

Variables	Mean	Median	Std	P25	P75	N
# Prior Machine Downloads	319.40	77.00	554.10	24.00	473.30	149,771
# Prior Other Downloads	92.88	35.25	272.40	15.75	83.75	149,771
Prior Machine Downloads	0.51	0.50	0.35	0.25	0.75	149,771
Prior Other Downloads	0.51	0.50	0.35	0.25	0.75	149,771
Machine Readability	-0.02	-0.04	1.48	-1.16	1.12	148,799
LM - Harvard	-2.48	-2.45	0.54	-2.81	-2.13	149,771
LM Sentiment	1.69	1.62	0.65	1.23	2.07	149,771
Harvard Sentiment	4.18	4.15	0.68	3.72	4.60	149,771
BM	0.67	0.56	1.48	0.31	0.89	142,198
Size	6.17	6.08	2.00	4.73	7.48	142,198
Momentum	0.15	0.07	0.71	-0.17	0.32	144,013
ROA	-0.01	0.01	0.39	0.00	0.02	148,402
# Institutional Owner	144.9	88	197.6	29	172	145,618
Product Market Fluidity	7.06	6.27	3.77	4.27	9.12	140,177
Investment-to-Assets	0.04	0.02	0.24	0.00	0.06	141,964
Book Leverage	4.49	1.95	183.80	1.40	3.50	142,324
Kaplan-Zingales Index	-50.47	-2.965	1,404	-12.1	0.402	139,656
Advertising Expense	0.02	0.00	0.10	0.00	0.01	142,324
R&D	0.04	0.00	0.11	0.00	0.04	142,324

Table 2. 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. See Table 1 for the variable definitions. The sample consists of 10-K and 10-Q filings between January 2004 and December 2015. *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)
Dependent Variable	*Machine Readability*	
Prior Machine Downloads	0.943***	0.946***
	(4.16)	(4.13)
Prior Other Downloads	-0.234	-0.234
	(-1.58)	(-1.56)
BM		0.003
		(1.10)
Size		-0.110***
		(-7.15)
Momentum		-0.071
		(-1.06)
ROA		0.004
		(0.52)
# Institutional Owner		0.000
		(1.31)
Product Market Fluidity		0.011**
		(2.27)
Investment-to-Assets		0.071***
		(3.71)
Book Leverage		-0.000**
		(-2.35)
Kaplan-Zingales Index		-0.000
		(-0.84)
Advertising Expense		-0.011
		(-0.23)
R&D		-0.100
		(-1.67)
N	153,152	139,019
Fixed Effects	Firm&Year	Firm&Year
Adjusted R^2	0.161	0.161

Table 3. Machine Downloads and Sentiment

This table reports the regression results examining the relation between the sentiment of a firm's filing and the machine downloads of the firm's past filings. See Table 1 for the variable definitions. The sample consists of 10-K and 10-Q filings between January 2004 and December 2015. *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*		*LM Sentiment*		*Harvard Sentiment*	
Prior Machine Downloads	-0.046***	-0.045***	-0.042***	-0.037***	0.004	0.008
	(-7.25)	(-7.00)	(-5.62)	(-4.97)	(0.54)	(1.04)
Prior Other Downloads	0.075***	0.080***	0.140***	0.139***	0.065***	0.059***
	(8.34)	(8.92)	(13.20)	(13.20)	(5.83)	(5.36)
BM		0.000		0.002		0.002
		(0.24)		(1.21)		(1.33)
Size		-0.014***		-0.079***		-0.065***
		(-2.60)		(-11.35)		(-9.50)
Momentum		0.014***		-0.014***		-0.028***
		(7.12)		(-5.41)		(-9.34)
ROA		-0.004*		-0.009		-0.005
		(-1.95)		(-1.55)		(-1.26)
# Institutional Owner		-0.000*		-0.000*		-0.000
		(-1.92)		(-1.84)		(-0.50)
Product Market Fluidity		0.009***		0.019***		0.010***
		(6.72)		(11.47)		(6.41)
Investment-to-Assets		0.015*		-0.068***		-0.083***
		(1.65)		(-3.46)		(-3.89)
Book Leverage		-0.000		0.000		0.000
		(-0.38)		(0.72)		(1.16)
Kaplan-Zingales Index		-0.000***		0.000***		0.000***
		(-2.64)		(4.65)		(12.86)
Advertising Expense		-0.012		0.024		0.037
		(-0.76)		(0.77)		(1.22)
R&D		0.016		-0.095**		-0.110***
		(0.70)		(-2.27)		(-3.82)
N	149,686	135,952	149,686	135,952	149,686	135,952
Fixed Effects	Firm&Year	Firm&Year	Firm&Year	Firm&Year	Firm&Year	Firm&Year
Adjusted R^2	0.542	0.542	0.606	0.610	0.582	0.584

Table 4. 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. *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). See Table 1 for the other variable definitions. The sample consists of 10-K and 10-Q filings between January 2004 and December 2015. *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*		*LM Sentiment*		*Harvard Sentiment*	
Prior Machine Downloads × Post	-0.130***	-0.120***	-0.161***	-0.147***	-0.031*	-0.026
	(-9.40)	(-8.74)	(-9.59)	(-8.99)	(-1.85)	(-1.60)
Prior Machine Downloads	0.018***	0.016**	0.059***	0.057***	0.041***	0.041***
	(2.63)	(2.39)	(7.28)	(7.30)	(4.98)	(5.11)
BM		0.000		0.002		0.002
		(0.06)		(1.10)		(1.28)
Size		-0.016***		-0.077***		-0.062***
		(-2.83)		(-10.82)		(-8.75)
Momentum		0.015***		-0.012***		-0.027***
		(7.20)		(-4.57)		(-8.71)
ROA		-0.004*		-0.009		-0.005
		(-1.88)		(-1.54)		(-1.28)
# Institutional Owner		-0.000		-0.000		-0.000
		(-0.87)		(-0.95)		(-0.36)
Product Market Fluidity		0.008***		0.017***		0.009***
		(6.01)		(10.42)		(5.75)
Investment-to-Assets		0.009		-0.073***		-0.082***
		(0.93)		(-3.58)		(-3.83)
Book Leverage		0.000		0.000		0.000
		(0.24)		(1.44)		(1.46)
Kaplan-Zingales Index		-0.000***		0.000***		0.000***
		(-2.59)		(6.53)		(14.13)
Advertising Expense		-0.010		0.029		0.039
		(-0.62)		(0.85)		(1.25)
R&D		0.012		-0.090**		-0.102***
		(0.49)		(-2.19)		(-3.79)
N	137,917	125,483	137,917	125,483	137,917	125,483
Fixed Effects	Firm&Year	Firm&Year	Firm&Year	Firm&Year	Firm&Year	Firm&Year
Adjusted R^2	0.537	0.537	0.604	0.608	0.580	0.582

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

This table reports the dynamic effect 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. *Between SSRN and Publish* in an indicator variable equal to one for filings in 2009 and 2010, and zero otherwise. *PublishYear* is an indicator variable equal to one for filings in 2011 and zero otherwise. *PostPublish_1Year* is an indicator variable equal to one for filings in 2012 and zero otherwise. *PostPublish_2Year+* is an indicator variable equal to one for filings in 2013 and onwards, and zero otherwise. *PreSSRN_1Year* is an indicator variable equal to one for filings in 2008 and zero otherwise. See Table 1 for the other variable definitions. The sample consists of 10-K and 10-Q filings between January 2004 and December 2015. *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.

Dependent Variable	*LM – Harvard*
Prior Machine Downloads × Between SSRN and Publish (2009 - 2010)	-0.104***
	(-7.38)
Prior Machine Downloads × PublishYear (2011)	-0.054***
	(-3.13)
Prior Machine Downloads × PostPublish_ 1Year (2012)	-0.128***
	(-7.46)
Prior Machine Downloads × PostPublish_2Year+ (2013 and later)	-0.150***
	(-9.20)
Prior Machine Downloads × PreSSRN_1Year (2008)	0.017
	(1.25)
Prior Machine Downloads	0.044***
	(5.26)
N	135,952
Firm-Level Controls	Yes
Fixed Effects	Firm&Year
Adjusted R^2	0.543

Table 6. Machine Downloads and Sentiment: Conditional on Past Machine Readability

This table examines the relation between the sentiment of a firm's filing and the machine downloads of the firm's past filings for subsamples partitioned based on the average *Machine Readability* of firms' historical filings. Each year, a firm's filing on day zero is sorted into two groups by the average *Machine Readability* of the firm's historical 10-K and 10-Q filings that were filed during days [-390, -30]. The group above (below) the median breakpoint is labeled as "High" ("Low"). See Table 1 for the variable definitions. The sample consists of 10-K and 10-Q filings between January 2004 and December 2015. *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, respectively, for the regression coefficients (two-tailed) and for the difference of coefficients (one-tailed).

Dependent Variable	*LM – Harvard*			
	Prior Machine Readability		Prior Machine Readability	
Subgroup	High	Low	High	Low
	(1)	(2)	(3)	(4)
Prior Machine Downloads × *Post*	-0.122***	-0.086***	-0.118***	-0.083***
	(-6.88)	(-4.60)	(-6.62)	(-4.44)
Prior Machine Downloads	0.007	-0.002	0.004	-0.001
	(0.80)	(-0.23)	(0.47)	(-0.14)
Diff. of coeff. (*Machine Downloads* × *Post*)	-0.036*		-0.035*	
p-value	0.081		0.088	
N	68,232	68,866	61,969	62,664
Firm-Level Controls	No	No	Yes	Yes
Fixed Effects	Firm&Year	Firm&Year	Firm&Year	Firm&Year
Adjusted R^2	0.570	0.570	0.569	0.571