Annual Review of Financial Economics

Textual Analysis in Finance

Tim Loughran and Bill McDonald

Mendoza College of Business, University of Notre Dame, Notre Dame, Indiana 46556-5646, USA; email: loughran.9@nd.edu

Annu. Rev. Financ. Econ. 2020. 12:357–75

The *Annual Review of Financial Economics* is online at financial.annualreviews.org

https://doi.org/10.1146/annurev-financial-012820-032249

JEL codes: D82, D83, G14, G18, G30, M40, M41

Keywords

textual analysis, complexity, machine learning, readability, Fog Index, social media, lexicons

Abstract

Textual analysis, implemented at scale, has become an important addition to the methodological toolbox of finance. In this review, given the proliferation of papers now using this method, we first provide an updated survey of the literature while focusing on a few broad topics—social media, political bias, and detecting fraud. We do not attempt to survey the various statistical methods and instead initially focus on the construction and use of lexicons in finance. We then center the discussion on readability as an attribute frequently incorporated in contemporaneous research, arguing that its use begs the question of what we are measuring. Finally, we discuss how the literature might build on the intent of measuring readability to measure something more appropriate and more broadly relevant—complexity.

1. INTRODUCTION

The rapidly changing nature of textual analysis in financial economics gives us the opportunity to review some significant papers from the last few years, beyond the prior literature reviews of Li (2010); Das (2014); Kearney & Liu (2014); Loughran & McDonald (2016); and Gentzkow, Kelly & Taddy (2019). We begin with a selective literature review that focuses on the use of textual analysis in social media, political bias, and fraud detection, as these are areas where we believe some recent papers have more broadly validated the usefulness of textual analysis. We then provide a discussion of the ongoing methodological debate concerning the formation of corpus-specific lexicons, essentially a discussion of humans versus machines.

After addressing these initial topics, we center our discussion on readability, a textual analysis measure that researchers continue using in the extant literature and one we argue is mis-specified. We believe, however, that careful consideration of what we actually measure with readability suggests a more encompassing and useful firm attribute that can be best proxied using textual analysis. That attribute is firm-level complexity.

2. TEXTUAL ANALYSIS AND SOCIAL MEDIA

Interestingly, one of the first applications of textual analysis in finance focused on social media taken from Yahoo stock forums (see Das & Chen 2007). More currently, one of the most novel sources of social media data is Twitter. Twitter is a microblogging social media service that started in 2006. Messages on Twitter (i.e., tweets) were initially limited to 140 characters, with the limit doubling to 280 in November of 2017. That tweets provide an instantaneous measure of new information is evidenced in geology, where tweets have been shown to provide immediate measures of earthquake intensity and better identify ShakeMaps (e.g., Burks, Miller & Zadeh 2014). In accounting, Blankespoor, Miller & White (2013) examine the role that Twitter plays in the information dissemination of tech firms. They find that companies using Twitter to send press-release links to investors are associated with greater abnormal depths and lower abnormal bid-ask spreads. In an experimental setting, Elliott, Grant & Hodge (2018) observe the role of trust that CEO Twitter accounts have in lessening the impact of negative news. Following negative firm news, they find that investors are more likely to buy shares of a firm whose CEO interacts with market participants via Twitter.

The Federal Reserve's Federal Open Market Committee (FOMC) holds eight regularly scheduled meetings each year to assess U.S. monetary policy. Creating a database of tweets discussing the FOMC, Azar & Lo (2016) examine the ability of social media to predict the returns of the Center for Research in Security Prices value-weighted stock index. Their trading strategy is quite simple. They use Topsy API—a service that indexes tweets—to target mentions of *FOMC*, *Federal Reserve*, *Bernanke* (during his chair term), or *Yellen* (during her chair term) and tabulate tweet sentiment using Pattern, a Python package. They score the polarity of each tweet between −1 and +1. Since individuals with more followers should have greater impact on index returns, the tweets are weighted by the number of the user's followers. Most of their analysis focuses on the 2009–2014 time period.

Azar & Lo (2016) find that the trailing sentiment of investor tweets about the Federal Reserve (Fed) has impact on value-weighted index returns. Higher tweet sentiment about the Fed is linked with higher subsequent index returns. The results are particularly strong on the eight FOMC dates each year. After controlling for common market factors, they find that on FOMC meeting dates, a one-standard-deviation increase in lagged tweet sentiment is associated with 0.62% higher returns.

Similarly, StockTwits, founded in 2008, is a social media platform that links investors and traders together and is essentially a focused application of Twitter. By 2019, StockTwits had two million registered members posting four million monthly messages about financial assets. Even though social media users represent only a small percentage of all traders, can sentiment from their Twitter and StockTwits messages affect market liquidity? Agrawal et al. (2018) report that social media messages are correlated with liquidity measures like turnover, mini-flash crash count, and the number of trades outside the quote spread at the intraday level. More bullish or bearish sentiment (defined as more than three standard deviations above or below the average for a particular stock) is associated with higher values for liquidity measures. Importantly, negative sentiment has a much bigger effect on liquidity than positive sentiment. The authors argue that panics should have a quicker impact on turnover or number of quotes than market upswings.

As an information source that is instantaneous and global, Twitter and StockTwits would seem to be gold mines of market-relevant data. One of the difficulties, however, in measuring sentiment from social media postings is the use of slang, profanity, symbols, and sarcasm. Standardized word lists like those from Loughran & McDonald (2011) might provide a foundational lexicon for capturing sentiment in this application, but clearly, they were not designed to extract sentiment from the dynamic language of tweets. The fluidity of words, symbols, and acronyms in tweets, unfortunately, makes the always-present signal-to-noise problem in textual analysis much more challenging.

Bartov, Faurel & Mohanram (2017) focus on tweets written in the nine trading days immediately before earning announcements. The authors use two different techniques to measure tweet sentiment. The first technique uses a naïve Bayes algorithm to categorize tweets into positive, negative, and neutral groups. The second technique uses three different dictionaries to gauge negative tweet sentiment: Loughran & McDonald's (2011) negative word list, the Harvard IV-4 TagNeg (H4N) word list, and the Hu & Liu (2004) word list created specifically for sentiment analysis in social media. Bartov, Faurel & Mohanram (2017) find that both techniques are positively associated with positive future realized earnings surprises and the instantaneous stock price reaction to the earnings announcements.

Constructing a crypto-specific lexicon, Chen et al. (2019) examine the sentiment of StockTwit tweets and their impact on cryptocurrency returns. They focus on an aggregate proxy of cryptocurrency value, the CRIX (CRyptocurrency IndeX). One positive aspect of StockTwits is that investors can label their message as being either bullish or bearish. Using more than one million different messages about cryptocurrencies, Chen et al. (2019) created a new lexicon that includes emojis, slang, and even profanity. They find that their word list is 32% better than the Loughran & McDonald (2011) lexicon when applied to an out-of-sample classification setting. Since the language of StockTwits includes intentional misspellings (*hodl!*—a misspelling of "hold," now said to imply "Hold On for Dear Life"), off-color language (*buttcoin*), and the use of rocket ship emojis to convey sentiment, they find that corpus-specific word lists dominate traditional lexicons like the H4N and Loughran & McDonald (2011) sentiment word lists (hereafter referred to as the Loughran–McDonald word lists), as would be expected.

Both of these papers provide good examples of the importance of adapting sentiment word lists to the corpus of interest. For example, the Loughran–McDonald sentiment word lists were developed specifically in the context of 10-K filings.[1] The application of their lexicon to tweets, earnings calls, or other sources requires that researchers carefully and transparently modify the

[1]The first public version of the Loughran–McDonald sentiment dictionaries appeared in Loughran & McDonald (2011). The lists and accompanying dictionary are updated biennially and made available at **https://sraf.nd.edu**.

foundational lists. The computational linguistics literature has long emphasized the importance of developing categories adapted to the corpus being studied (see Berelson 1952).

3. TEXTUAL ANALYSIS AND PARTISAN SLANT

Extending the earlier political slant work by Gentzkow & Shapiro (2010), Gentzkow, Shapiro & Taddy (2019) use a lasso-type estimator to analyze partisan language spoken in the U.S. Congress. To measure the magnitude of partisan differences in speech, the three authors examine all unique phrases (508,352 in total) used by 7,732 unique speakers in the U.S. Congress from 1873 to 2016. Their language-based method addresses the severe finite-sample bias present in other standard approaches. Gentzkow, Shapiro & Taddy (2019) document that partisan language is much more common in recent years than in the past.

Using the Gentzkow, Shapiro & Taddy (2019) methodology, Engelberg et al. (2019) examine the presence of partisan language in the speeches of commissioners from the Securities and Exchange Commission (SEC) and members of the Fed Board of Governors during the 1930–2016 time period. As might be expected, in recent decades Republican SEC Commissioners use phrases like "unintended consequences of regulation" much more often than their Democratic counterparts do. Conversely, speeches by Democratic SEC Commissioners disproportionally mention "board diversity." Their Congress-based regulator partisanship measure finds that SEC Commissioners showed an increasing level of partisan language starting in the mid-1970s. The Fed Governors, conversely, showed no upward trend in partisan language compared with members of Congress.

4. DETECTING FRAUD

One of the holy grails of business research is to identify a discriminating and robust way to flag fraudulent activity by company insiders. Although establishing a time frame for identifying fraud is elusive, the SEC issues Accounting and Auditing Enforcement Releases (AAERs) during or after an investigation alleging accounting fraud, which allows researchers to compare company characteristics during the time period of the alleged misreporting with other contemporary firms. Typically, researchers find significant differences between the AAER and non-AAER samples. We next discuss several examples of recent papers addressing this important topic.

Much of the prior research in this area focuses on quantitative accounting numbers to identify fraud by managers. As an example, Dechow et al. (2011) link companies with unusually high levels of off-balance-sheet items like operating leases or low accrual quality with higher misstated quarterly or annual earnings. Textual analysis allows researchers to determine if it is possible to identify fraudulent activity simply by managers' choice of words or topics in their annual reports.

Hoberg & Lewis (2017) examine, in the Management Discussion and Analysis (MD&A) section of a 10-K, whether word usage differs between AAER firms and industry-age-size matched non-AAER firms. They create a fraud score variable that is fitted using in-sample filings from 1997 to 2001 and use the 2002–2010 time period as an out-of-sample test. Firms that exhibit similarity with the abnormal vocabulary associated with AAER companies have a significantly higher probability of ex post accounting misstatements. That is, specific vocabulary choices can actually be used to help predict out-of-sample accounting fraud.

Hoberg & Lewis (2017) also use latent Dirichlet allocation (LDA) to examine differences in topics selected by AAER and non-AAER firms. LDA is a generative, unsupervised method for identifying latent attributes—essentially cluster analysis for words—producing "topics," i.e., word groups with common context. AAER companies tend to have abnormally long discussions of

acquisitions, hedging transactions, derivative instruments, and business opportunities. Topics in the MD&A section that are significantly underdisclosed by AAER firms include realized gains, marketing expenses, professional fees, legal proceedings, and research development. Interestingly, they find that managers tend to disassociate their personal names from fraudulent and irregular activities. AAER firms less frequently discuss LDA-identified topics that link the CEO with participation in actual firm plans and financial strategies. Thus, managers seem to disassociate their names and titles in the MD&A section when they are committing fraud. Hoberg & Lewis (2017) argue that the avoidance of manager names is "likely to insulate themselves from fallout should the fraud be discovered in the future" (p. 77).

Using accounting irregularity samples from AAERs, audit analytics, and amended 10-K filings, Brown, Crowley & Elliott (2020) use LDA to identify topics within the entire 10-K filing that are linked with financial misreporting. Their LDA algorithm is used over a rolling 5-year window to account for the changing nature of topics and language usage. Examples of topics include changes in income performance, postretirement benefit assumptions, and real estate loan operations. They find that semantically meaningful topics can assist researchers and investors in identifying out-of-sample firm misreporting even after controlling for financial information, document tone, document length, and other writing style characteristics.

Can innocent employees unknowingly transcribe fraudulent language into annual reports? In an experimental setting, Murphy, Purda & Skillicorn (2018) test whether individuals can unintentionally transfer deceptive words into the MD&A section of the Form 10-K. The researchers' baseline experiment starts with a CFO instructional memo for the preparation of the MD&A section. The memo contains actual language from an annual report that was fraudulent. Some of the participants were given a memo containing added Loughran & McDonald (2011) negative words (i.e., *argue*, *concern*, *diminished*, and *impair*) that have been associated with fraud. Murphy, Purda & Skillicorn (2018) find that participants unknowingly transfer linguistic cues from the CFO memo into the MD&A text they author.

Early studies in finance focused on linking linguistic sentiment in the media, such as the news, SEC filings, earnings conference calls, or tweets, with stock returns. The studies we have highlighted show that textual analysis is also useful in other classification and predictive tasks important in finance, even after controlling for traditional quantitative variables.

5. BUILDING LEXICONS: HUMANS VERSUS MACHINES

Although textual analysis dates back centuries (see Loughran & McDonald 2016), its explosive growth over the past decade is attributable to exponential growth in computational power and Internet content. Textual analysis is computationally intensive, so increasing computational power makes it more accessible, but of equal importance is the explosion of unstructured data made available through online repositories and social platforms. As with any growth area in research, a plethora of more sophisticated and hopefully more discerning methods are being adapted and developed; however, among the central tools in the textual analysis toolbox, simple bag-of-words methods and word lists continue to be used in assessing document sentiment or tone.

In finance and accounting, one of the first word lists gauging sentiment in business documents was by Loughran & McDonald (2011). They created six different word lists (negative, positive, uncertainty, litigious, strong modal, and weak modal word categories) specifically designed for the language used in business disclosures. In tone analysis, the focus is usually on negative sentiment. Most, but not all, research finds that investors tend to focus on pessimistic language in annual reports and newspaper articles while giving less attention to positive words (see Tetlock 2007; Loughran & McDonald 2011).

For alternative sources, such as earnings conference calls or press releases, however, both positive and negative word frequencies are typically controlled for (see Mayew & Venkatachalam 2012; Froot et al. 2017; Burks et al. 2018; Chen, Nagar & Schoenfeld 2018; Edmans et al. 2018). Positive words are less straightforward in their contextual meaning, which is why their relevance is somewhat dependent on the medium. In business reporting, negative words are rarely used unless absolutely necessary. Positive words in mandated financial disclosures, however, are frequently used to lessen the impact of the negative words necessary to describe a financial outcome; thus, their effective sentiment can be ambiguous.

Prior to the publication of the Loughran & McDonald (2011) word lists, the literature typically used sentiment dictionaries created by the psychology and sociology fields to measure the tone of business documents (i.e., the Harvard Psychosociological Dictionary). However, the negative-sentiment H4N word list has significant limitations when applied to business disclosures. Loughran & McDonald (2011) document that almost 75% of negative word counts of the Harvard Dictionary are misclassified. Commonly occurring Harvard negative words like *tax*, *excess*, *capital*, *board*, *foreign*, and *liabilities* are clearly not negative when used in financial disclosures.

Loughran & McDonald (2011) created their negative word list by deciding whether a particular word most likely has a negative meaning when used in a financial setting. They report that the most commonly appearing negative words in U.S. annual reports are *loss*, *losses*, *claims*, *impairment*, *against*, and *adverse*. Although much of the textual analysis literature has focused on firm-level sentiment, Jiang et al. (2019) take a completely different approach by creating an aggregate measure of sentiment from both mandatory and voluntary disclosures by companies. Using the Loughran & McDonald (2011) positive and negative word lists, the authors generate monthly sentiment from 10-Ks, 10-Qs, and earnings conference calls. Interestingly, the more positive the monthly sentiment index is, the lower subsequent market stock returns will be. Thus, their manager sentiment index is a contrarian stock market predictor.

Quite a number of alternative investor/consumer sentiment indexes are widely used in the literature [see, for example, Baker & Wurgler (2006) or the University of Michigan Consumer Sentiment Index (**http://www.sca.isr.umich.edu/**)]. Jiang et al. (2019) find that their monthly manager sentiment index is independent of the other sentiment measures. For example, although their manager sentiment index has a 0.53 correlation with the Baker & Wurgler (2006) investor sentiment index, in regressions with monthly excess returns as the dependent variable, their sentiment index has significantly higher R-squared values (9.75% versus 5.11%). They find that management sentiment is positively linked with overinvestment by insiders.

A number of other word lists have been created in the literature during the last decade. In an attempt to identify deceptive behavior from managers during earnings conference calls, Larcker & Zakolyukina (2012) created extreme negative and positive word lists. They find that deceptive CEOs more frequently use extreme positive language (e.g., *fabulous*, *marvelous*, and *peachy*) in the conference calls. Hope & Wang (2018), using the Larcker & Zakolyukina (2012) methodology, report that firms' bid-ask spreads significantly increase immediately following an accounting bigbath write-off by deceptive CEOs. This is evidence that investors can infer managerial deception from earnings conference calls.

To measure the level of financial constraints of publicly traded companies, Bodnaruk, Loughran & McDonald (2015) created a list of 184 words. They developed the word list by examining words contained in at least 5% of all annual reports and selecting tokens that would typically be considered constraining by the reader. The five most commonly occurring words from their list are *required*, *obligations*, *requirements*, *require*, and *impairment*. The authors find that their measure predicts subsequent financial outcomes like dividend omissions, equity recycling, and underfunded pensions better than other constraint indexes based on accounting variables. Soo (2018) created a

housing market sentiment index for 34 U.S. cities during the 2000–2013 time period on the basis of media articles. Her housing market sentiment index positively predicts future housing price growth over the subsequent 2 years. That is, more newspaper housing articles in a city containing tokens like *highs*, *frenziness*, *record*, and *booming* are associated with higher subsequent housing prices.

To examine the ability of traders to incorporate news into asset prices, Loughran, McDonald & Pragidis (2019) created a list of 130 keywords that should affect oil prices. They find significant short-term overreaction to oil-related news coverage. Examples of their keywords include *recovery*, *problems*, *attacks*, *oversupply*, and *hurricane*. All of the word lists from these studies were created by having people with at least some expertise examine conference calls, annual reports, or news articles.

Using a machine learning technique, support vector regression (SVR), Manela & Moreira (2017) create an uncertainty measure from front-page articles of the *Wall Street Journal* (WSJ).[2] They identify comovement between word frequencies in WSJ articles and the options implied volatility (VIX) using an SVR during 1996–2009. This is their training period. The 1986–1995 period is used as the out-of-sample test of their model fit. Then the authors have a prediction subperiod when the VIX is not available (i.e., 1890–1985). Their uncertainty measure is called news implied volatility (NVIX).

Manela & Moreira (2017) find that the higher the NVIX Index is (implying high uncertainty language in the WSJ articles), the higher subsequent market stock returns will be. The results are economically important; a one-standard-deviation increase in NVIX is associated with higher annual returns of 3.3% the following year. The time series pattern of NVIX produces mixed results. The NVIX correctly shows peaks at the 1929 stock market crash, the 2008 financial crisis, and the 1998 Long-Term Capital Management collapse. However, the NVIX is surprisingly low at the news of the sinking of much of the U.S. Pacific Fleet at Pearl Harbor (December 1941), the 1973 oil embargo crisis, and the bursting of the Internet bubble in 2000.

In contrast to the machine learning techniques, Baker, Bloom & Davis (2016) use humans to create their economic policy uncertainty (EPU) index. The strength of their paper is the simplistic tabulation of their uncertainty index. It is a count of articles in 10 major U.S. newspapers containing the trio of terms *economic* or *economy*; *uncertain* or *uncertainty*; and at least one of *Congress*, *deficit*, *Federal Reserve*, *legislation*, *regulation*, or *White House*. As might be expected, their EPU index spikes at the 2011 debt-ceiling dispute, the 9/11 attacks, the Lehman Brothers collapse, and both Gulf Wars. Higher levels of their EPU index are associated with higher stock return volatility and lower levels of investment at the firm level.

The three authors also develop a measure of health care and national security uncertainty from popular press articles. For health care uncertainty, words like *hospital*, *health insurance*, and *health care* are tabulated from newspaper articles, while tokens like *war* and *terrorism* are included in the national security uncertainty index. During and after the vote on the Affordable Care Act (i.e., Obamacare), the health care uncertainty index was at an elevated level.

Expanding on the prior literature, Bybee et al. (2019) gauge the state of the economy by analyzing the topics appearing in WSJ articles. Unlike Manela & Moreira (2017), who only examine front-page WSJ articles, or Baker, Bloom & Davis (2016), who do a key word search from a limited number of newspapers, Bybee et al. (2019) examine all WSJ articles—763,887 separate articles in total—during 1984–2017. Their paper uses LDA, an unsupervised modeling approach, to generate the topics. The four authors find a link between topic coverage and the economy. A one-standard-deviation increase in recession news is associated with a 1.7% decline in the

[2] See Gentzkow, Kelly & Taddy (2019) section 3.1.3 for a discussion of the SVR method.

following year's industrial production. As expected, frequency of articles in the WSJ on the topic of "terrorism" spiked immediately after 9/11 and remained at elevated levels for the following years.

Are humans or computers better at identifying specific words that are useful in gauging the tone of a large corpus of financial disclosures? Many of the prior papers use subjectively determined word lists to test for sentiment. Researchers hesitate to define a word list because of this subjectivity. For this approach to be effective, the process must be transparent, and the resulting lists should be reasonably exhaustive. Making the lists exhaustive precludes the potential for p-hacking a list down to the most ex post powerful words or having managers simply avoid identified words in crafting their future documents.

Some would argue that using computational methods to derive sentiment word lists avoids the subjectivity of expert selection. Most of the recently developed methods are taken from the machine learning field, typically falling under the labels of supervised or unsupervised learning. Supervised learning is used in instances where some arbiter of truth is available, e.g., the firm did or did not go bankrupt or the magnitude of a stock return. In unsupervised methods, researchers allow the technique to look for hidden structure in the data, e.g., topic analysis. To avoid overfitting, a holdout sample is used to develop a model that can then be applied to out-of-sample data.

Essentially, the challenge in using computational methods is one of dimension reduction. Using only single words, we might have more than 80,000 features attempting to predict the dependent variable. If we expand this to phrases (n-grams), the problem grows exponentially. Many existing techniques taken from the computational linguistics literature—e.g., latent semantic analysis, LDA, topic analysis—have been applied in finance and accounting applications and are reviewed in Gentzkow, Kelly & Taddy (2019). We focus on a new and promising technique developed in a recent working paper by Ke, Kelly & Xiu (2019).

Ke, Kelly & Xiu (2019) apply a new supervised machine learning approach on a long time series of *Dow Jones Newswire* articles. Instead of using a predefined dictionary to define document sentiment, the authors "learn the sentiment scoring model from the joint behavior of article text and stock returns" (Ke, Kelly & Xiu 2019, p. 4). Their model is both transparent and tractable. In a daily, equally weighted trading strategy where the top 50 stocks in terms of positive sentiment are purchased while the 50 stocks with the lowest sentiment are shorted, the realized annualized Sharpe ratio is 4.29.

Typical of machine learning, rolling session windows are used to train the model. The authors use rolling 15-year sessions—the first 10 years to train the model and the last 5 years to validate the model. Since the training sessions use different news articles, the sentiment word lists can fluctuate. However, these nine tokens are consistently on the negative word list across all their 14 training sessions: *shortfall*, *downgrade*, *disappointing*, *auditor*, *tumble*, *blame*, *hurt*, *plunge*, and *slowdown*.

Of the nine words, only four are not on the Loughran & McDonald (2011) negative word list (*auditor*, *tumble*, *blame*, and *plunge*). Recall that the Loughran–McDonald word lists were created using text from annual reports. The word *auditor* is a common token in the annual report that does not necessarily have negative meaning. That is, every firm mentions their auditor within the Form 10-K. Yet, if a news article uses the word *auditor* when discussing a company, this is typically not good news.

Since the tokens *tumble*, *blame*, and *plunge* are clearly strongly pessimistic words, why are they not in the Loughran–McDonald negative dictionary? The three words in question are extreme emotion terms that infrequently appear in annual reports and thus would not be picked up by a screening mechanism that looks for words appearing in at least 5% of all 10-Ks. For example, according to the 2018 Loughran–McDonald Master Dictionary, *tumble* appeared in U.S.

annual reports over the last few decades only 371 times compared with a 140,029 count for *slowdown*.

Gentzkow, Kelly & Taddy (2019) in their review of "text as data" focus more on methods and note that "a large share of text analysis applications continue to rely on ad hoc dictionary methods rather than deploying more sophisticated methods for feature selection" (p. 569). Although they note that dictionary methods might be best in some instances, they argue that ultimately "modern methods" from machine learning will win in terms of performance. From our experience in using both approaches, we are less optimistic about future solutions and have found the dictionary methods much less likely to capture data artifacts.

One of the central puzzles in finance provides an interesting parallel with this issue in textual analysis. The crux of asset pricing, much like textual analysis, is one of dimensionality reduction. Can we, from thousands of securities, identify a few meaningful common factors? Although the finance literature has experimented with statistical data reduction (e.g., principal component analysis or factor analysis), the dominant factor model in finance is based on Fama & French's (1993) linking of firm characteristics that exhibit empirical regularities with stock returns.

Another concern in comparing machine learning methods with sentiment lexicons is whether the fundamental hypotheses are the same. Without question, a machine learning method should be able to identify a collection of words that predict, for example, stock returns better than a fixed sentiment lexicon. Are these words, however, in fact capturing sentiment per se or might they be identifying firm attributes that happen to produce positive (or negative) outcomes both in-sample and out-of-sample?

Although approaches such as those proposed by Ke, Kelly & Xiu (2019) provide a promising direction for computational selection of word lists, in our experience such lists include too many words that are simply pseudodummy variables identifying a particular firm or industry with outcome measures of large magnitude, and these methods' out-of-sample characteristics are fragile. Of the words listed in their top 50 positive and negative words, many appear idiosyncratic. In addition, an inexhaustive selection of words creates an endogeneity problem going forward, i.e., if a small vocabulary of negative words is identified, managers will in the future avoid them. Ultimately, the empirical success of these different approaches will impact the choice of methods going forward.

6. READABILITY

Our greatest concern in the evolution of textual methods in the finance and accounting disciplines is the continued focus on readability, typically as a control, when examining other economic relations. In this segment, we discuss the literature surrounding readability in financial documents before proposing the measurement of a broader attribute that is amenable to textual solutions.

We first note that a precursor to the importance of measuring readability is the assumption that investors actually read the filings. Historically, accounting researchers have shown minimal market reactions to the filing of quarterly or annual reports (see Griffin 2003). Loughran & McDonald (2017) show that the average publicly traded firm's 10-K is downloaded only approximately 28 times immediately after the filing. In addition, Cohen, Malloy & Nguyen (2020) show that investors are slow in incorporating 10-K information into stock prices. They argue that quarterly and annual reports contain significant amounts of valuable information that investors appear to ignore. To prove their point, the three authors document that a trading portfolio going long on "nonchangers" and short on "changers" generates large abnormal returns (up to 188 basis points in monthly alphas). Their evidence is consistent with the notion of inattention by investors to simple changes over time in public disclosures.

6.1. Problems with Using the Fog Index as a Measure of Readability

Fog Index: equal to 0.4 * (average number of words per sentence + fraction of complex words), where complex words are those with more than two syllables

If we are trying to characterize the information environment of a firm, especially when basing a sample on financial filings, an obvious prima facie choice is to measure the readability of the disclosures. Quantifying the readability of financial text is a challenging endeavor that regulators and academics have been struggling with for decades. One of the most influential papers on the topic of readability in the accounting literature is Li (2008). Using the Fog Index—a readability metric originally designed in the early 1950s to differentiate grade school reading material—Li (2008) gauges the readability of U.S. annual reports and finds that less readable annual reports (i.e., those with higher Fog Index values) are linked with lower earnings.

The attraction of the Fog Index for researchers is its easy tabulation and its simplistic output of the number of years of formal education needed to understand the document in a first reading. That is, a Fog Index value of 19 (i.e., a typical value for U.S. annual reports) implies that the reader will need more than an MBA in terms of formal schooling to understand the document in an initial reading. The formula for the Fog Index is:

$$\text{Fog Index} = 0.4^{*}\ (\text{average number of words per sentence} + \text{fraction of complex words}), \quad 1.$$

where complex words are defined as words with more than two syllables and higher values of the Fog Index imply less readable text.

As pointed out by Loughran & McDonald (2014a), the Fog Index is a poor measure of business document readability for several reasons. First, the fraction of complex words is a flawed metric since the tokens most frequently appearing in business documents with more than two syllables are typically trivial for investors to comprehend. Loughran & McDonald (2014a) report that the most frequently appearing complex words in annual reports are *financial*, *company*, *interest*, *agreement*, *including*, *operating*, *period*, and *related*. The most commonly appearing word with the largest number of syllables is *telecommunications*, hardly a word requiring dictionary sourcing. Second, it is difficult to calculate correctly the average number of words per sentence in a complex document like an annual report.[3] Empirically, Loughran & McDonald (2014a) show that a number of different measures of readability—the natural log of the text document 10-K file size in megabytes, commonality of words, count of jargon words, and number of words in the document—all perform significantly better than the Fog Index when the corpus is business disclosures.

Yet even after the sharp criticism by Loughran & McDonald (2014a), the accounting and finance literature continues to use the Fog Index as a measure of financial disclosure readability. For example, Lo, Ramos & Rogo (2017) use the Fog Index of the annual report's MD&A section as their dependent variable. They find that companies managing their earnings to beat the prior year's benchmark value have an MD&A section with higher Fog Index scores.

In examining trends in annual report disclosure attributes and topics over time, Dyer, Lang & Stice-Lawrence (2017) use the Fog Index as their readability measure. As others have noted, the three authors mention the decline in readability of annual reports over the recent decades (i.e., higher Fog Index values and longer length). Using LDA, they find that only three topics (internal controls, fair value, and risk factor disclosures) account for the vast majority of the increasing length of U.S. annual reports.

In creating a new measure of accounting complexity (a simple count of XBRL accounting tags), Hoitash & Hoitash (2018) use the Fog Index as a measure of linguistic complexity of annual reports. Although one might expect a positive correlation between financial accounting complexity

[3] See the criticism by Bushee, Gow & Taylor (2018) of Li's (2008) use of the Lingua::EN::Fathom Perl routine to calculate the average number of words per sentence. This particular Perl routine systematically understates the correct Fog Index value when used in complicated documents like annual reports.

and linguistic complexity, Hoitash & Hoitash (2018) find that counts of XBRL tags in the Form 10-K and the Fog Index are negatively correlated.

The Fog Index of annual reports is positively associated with the board of director's size and the fraction of the board with accounting expertise, as reported in a paper by Chychyla, Leone & Minutti-Meza (2019). The authors suggest that the Fog Index is a proxy for financial reporting complexity. Firms with more financial complexity should have more board of director members and a higher fraction of directors with accounting expertise.

Flesch–Kincaid Index: equal to 0.39 * (number of words / number of sentences) + 11.8 * (number of syllables / number of words) − 15.59

Using a sample of 1,581 bilateral strategic alliances during 1995–2012, Baxamusa, Jalal & Jha (2018) examine the market's reaction to the announcement of an alliance. They find that short-term announcement returns are lower when the firm's partner has higher Fog Index values for their annual reports. The authors argue that when the partner firm uses longer sentences or a higher fraction of words with more than two syllables in length (i.e., higher Fog Index values), the market views the firm as less credible and thus the partnership has a lower probability of being successful. Chakrabarty et al. (2018) examine the linkage between managers' risk-taking behavior and the readability of the annual report. They find that when managers are awarded with high vega compensation (i.e., high dollar change in the CEO's stock option portfolio for a 1% change in the firm's stock return volatility), the company issues less readable subsequent annual reports. In a robustness section of their paper, they use the Fog Index as an alternative measure of Form 10-K readability.

Bushee, Gow & Taylor (2018) use the Fog Index as their paper's measure of linguistic complexity in the context of quarterly conference calls. The three authors attempt to separate linguistic complexity into two different components, obfuscation and information. They obtain a measure of the information component by running a regression with the Fog Index of insider's conference call text as the dependent variable. The Fog Index of the analyst's question is the independent variable. The obfuscation component is the residual from the regression. In both the presentation and discussion sections of the earnings conference call, they find a negative linkage between the estimated information component and information asymmetry, and a positive association between obfuscation and information asymmetry.

6.2. Strong Correlations Among Various Established Readability Measures

To allay concerns based on the results of Loughran & McDonald (2014a) or as a robustness test, authors sometimes use a variety of alternative readability measures. For example, in a footnote, Li (2008) notes that if the Fog Index is replaced with the Flesch–Kincaid Index as a proxy for readability, his results are similar.[4] Brown, Crowley & Elliott (2020) use both the Fog Index and the Coleman–Liau Index as complementary measures of readability in their attempt to identify company financial misreporting from annual report text. Smales & Apergis (2017) use the Flesch–Kincaid score to measure the linguistic complexity of the FOMC decision statement, whereas Hayo, Henseler & Rapp (2019) use the Flesch–Kincaid Index as a measure of verbal complexity for the introductory statements from the European Central Bank's Governing Council press conferences.

Instead of using only one readability measure that might be influenced by measurement error, Guay, Samuels & Taylor (2016) combine six different indexes [Fog, Flesch–Kincaid, LIX (Läsbarhetsindex), RIX (Rate Index), ARI (Automated Readability Index), and SMOG (Simple

[4]Flesch–Kincaid offers two methods, the reading-ease formula and the grade-level formula, which are simply linear transforms of average-words-per-sentence along with average-syllables-per-word. The latter is a modification from the second term of the Fog Index but, as we show later in **Table 1**, they are highly correlated.

Table 1 Correlations between various readability indexes

	Fog	Flesch–Kincaid	LIX	RIX	ARI	SMOG	Log(file size)
Flesch–Kincaid	0.96	NA	NA	NA	NA	NA	NA
LIX	0.93	0.94	NA	NA	NA	NA	NA
RIX	0.95	0.96	0.99	NA	NA	NA	NA
ARI	0.89	0.91	0.92	0.94	NA	NA	NA
SMOG	0.97	0.93	0.90	0.93	0.84	NA	NA
Log(file size)	0.23	0.31	0.22	0.23	0.11	0.25	NA
Log(word count)	0.34	0.44	0.37	0.39	0.31	0.36	0.68

This table reports the correlations between eight different readability indexes using all firms filing a Form 10-K (i.e., annual report) on EDGAR (Electronic Data Gathering, Analysis, and Retrieval) containing at least 3,000 words between January 1997 and July 2019. The data are obtained from Wharton Research Data Services. The number of firm-year observations is 165,079. The Fog, Flesch–Kincaid, LIX (Läsbarhetsindex), RIX (Rate Index), ARI (Automated Readability Index), and SMOG (Simple Measure of Gobbledygook) Indexes are created using a differing combination of document words per sentence, fraction of complex words, and syllable counts per word. Log(file size) is the natural log of the annual report document file size from EDGAR in megabytes. Log(word count) is the natural log of the number of words in the annual report.
Other abbreviation: NA, not applicable.

Measure of Gobbledygook)] to create their ReadIndex variable. The correlations between these readability measures are extremely high because the indexes are very similar in nature, making the diversity of measures less effective at demonstrating robustness.

Table 1 reports the correlations between widely used readability measures for a large sample of annual reports (i.e., Form 10-K) obtained from Wharton Research Data Services for the time period between January 1997 and July 2019. Both public and private firms filing annual reports on the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website are included in the sample. The only data screen is that the annual report must contain at least 3,000 words. The eight different readability measures are the Fog Index, Flesch–Kincaid Index, LIX Index, RIX Index, ARI Index, SMOG Index, Log(file size), and Log(word count). The total number of firm-year observations is 165,079.

The correlations between the Fog, Flesch, LIX, RIX, ARI, and SMOG Indexes are quite high. For example, the correlation between the Fog Index and Flesch Index is 0.96, whereas the value is 0.99 between the LIX and RIX Indexes. Given that some of the readability measure definitions differ only slightly, this should not be a surprise. That is, the LIX Index is average number of words per sentence plus the fraction of words over six characters in length, whereas the RIX Index is a count of words more than seven characters in length divided by the number of sentences. Since Fog, Flesch, and ARI assign a U.S. grade level for the readability of the document using average words per sentence and either the fraction of complex words (Fog) or the average of characters per word (Flesch and ARI), their values are similar in nature. The simplicity of many readability measures is attributable to the metrics being created before the widespread use of computers. Flesch was created in 1948, Fog in 1952, and ARI in 1967.

Although the correlations between the six traditional readability measures and Log(file size) and Log(word count) are positive, the values are much lower than the correlations with each other. For example, the Fog Index has a correlation of 0.23 with Log(file size) and 0.34 with Log(word count). The highest correlation Log(word count) has with the other measures is 0.68 with Log(file size). As the number of words in the annual report increases, so should the total document size.

It is important to note that the Form 10-K file size measure includes much more than words. Annual report file size also incorporates pictures, tables, graphics, and hypertext markup language (HTML) code. The **Table 1** correlations present some suggestive evidence that Log(file size) and Log(word count) offer alternative measures of readability for the literature beyond traditional

LIX Index (Läsbarhetsindex): defined as (number of words / number of sentences) + (number of words over 6 characters * 100) / number of words

RIX Index (Rate Index): equal to (number of words of length of 7 characters or more) / (number of sentences)

ARI Index (Automated Readability Index): equal to 4.71(number of characters / number of words) + 0.5(number of words / number of sentences) − 21.43

SMOG Index (Simple Measure of Gobbledygook): equal to 1.043 * sqrt(number of complex words * 30 / number of sentences) + 3.1291

metrics that were created decades ago and differ only slightly in how the metrics are created. In Loughran & McDonald (2014a), they recommend using the log of gross file size as a rough proxy to readability and show that it is most effective in measuring outcomes, such as earnings response coefficients, that might be associated with readability.[5]

Log(file size): the natural log of the annual report document file size from EDGAR in megabytes

Log(word count): the natural log of the number of words contained in the annual report (i.e., Form 10-K)

6.3. Possible Methodology to Salvage the Fog Index

Can a simple alteration in the tabulation of the Fog Index improve its ability to gauge the readability of financial documents? Kim, Wang & Zhang (2019) propose a fix to the criticism leveled by Loughran & McDonald (2014a) on the Fog Index concerning multisyllable words that are easily understood by investors. The three authors manually collect 2,028 words more than two syllables in length from the Compustat variable list and the Fama & French (1997) 49-industry description file to identify common business terminology that investors would easily comprehend. In the tabulation of the modified Fog Index, these common multisyllable words are coded as two-syllable words, thereby dramatically lowering their typical Fog Index value. Their average modified Fog Index is 12.96, compared with the average raw Fog Index of 19.69. Examples of the multisyllable words they identify as being well-known to investors include *acquisition*, *auditor*, *derivatives*, *intangibles*, *personnel*, *purchasing*, and *unconsolidated*.

As a robustness check, Kim, Wang & Zhang (2019) validate that their modified Fog Index value is statistically significant when post-filing-date return volatility is the dependent variable in the presence of 10-K file size. When the raw Fog Index is an independent variable, its significance completely disappears in the presence of the natural logarithm of the 10-K file size in megabytes. Thus, the authors have improved the applicability of the Fog Index when applied to business disclosures; however, as we discuss later they have not overcome the essential criticism of measuring readability.

6.4. Bog Index

Given the literature's use of dated, simplistic readability measures like the Fog Index and the Flesch Index, what other readability measures might provide a better alternative? Bonsall et al. (2017) introduce the Bog Index, which is generated from a proprietary software, StyleWriter—The Plain English Editor (**https://editorsoftware.com/stylewriter.html**). The advantages of using a proprietary software to gauge the readability of text is that the program attempts to incorporate the document's complex word count, sentence length, use of passive voice, weak verbs, and even jargon into its score. The measure tries to assess the writing quality of the document.

For instance, instead of using syllable count to define complex words, the StyleWriter software uses a proprietary weighting scheme of more than 200,000 words. Their complex word scale ranges from 0 (familiar) to 4 (abstract). Thus, tokens like *company*, *financial*, and *interest* have a Bog word score of 0, whereas words like *operate*, *approximately*, *acquisition*, and *generally* are given a midrange difficulty score of 2. Examples of words the StyleWriter software considers abstract (i.e., Bog word score of 4) listed in appendix D of Bonsall et al. (2017) include *alpinist* (a climber of high mountains), *Archaean* (a proper noun relating to the geologic age from approximately 3,800 to 2,500 million years ago), and *arioso* (a type of solo vocal piece occurring in an opera).

[5] Although some assumed that Loughran & McDonald (2014a) overlooked the impact of items such as pictures embedded as ASCII, which take up a substantial amount of space, they simply argued for gross file size because it is far easier to measure than net file size (where extraneous text is removed) and the two are very highly correlated.

How often do managers mention Alpine mountain climbers, the Archaean age, or an opera solo in their annual reports? Using EDGAR's full-text search in thousands of 10-K filings over the last 4 years, we find that *alpinist* never appears, *Archaean* occurs only seven times (by three different mining companies), and *arioso* appears a total of nine times by a single company always referring to one of their filtration product lines (Arioso Membrane Composite). Since the Loughran–McDonald Master Dictionary does not include abbreviations, acronyms, or proper nouns, the total word count for these three words in thousands of 10-Ks over the last 4 years would be zero.

The fact that the StyleWriter software individually scores more than 200,000 words based on familiarity is intriguing at first pass. However, as noted by Loughran & McDonald (2014a), most business disclosures do not differ dramatically in their use of polysyllabic words. As mentioned above, the StyleWriter software surprisingly codes the tokens *approximately* and *generally* as being mid-tier familiarity for readers. Yet, these two words appear, respectively, 19.6 million and 11.4 million total times in annual reports over the last few decades according to the 2018 Loughran–McDonald Master Dictionary. They are some of the most commonly occurring multisyllable words in financial disclosures. No typical reader of an annual report is going to stumble over words like *approximately* or *generally*.

Using proprietary software like StyleWriter to gauge the readability of financial disclosures is problematic. First, as much as the Fog Index is mis-specified in measuring business document readability, at least the Fog Index is completely transparent. Other researchers can easily replicate the Fog Index. Obviously, other researchers could replicate the results of Bonsall et al. (2017) if the software is purchased; however, given the proprietary nature of the 200,000 Bog word scores or the actual technique to identify the jargon/passive voice, researchers do not know exactly how the Bog Index is being created. Second, and more importantly, the Bog Index is a measure of writing style, which is not necessarily the same as readability.

6.5. Measuring Readability via Style

In August 1998, the SEC released *A Plain English Handbook* (Off. Invest. Educ. Assist. 1998). The stated purpose of the handbook was to provide the reader with helpful suggestions to create plain English financial disclosures. Some recommendations of Rule 421(d) include keeping sentences short, using the active voice, and eliminating legal jargon. Loughran & McDonald (2014b) show that the mandate had measurable impacts on Form 424s, IPO prospectuses, and 10-K filings. One simple way to measure the impact of this initiative is to count the use of personal pronouns over time, which in the SEC's document is recommended to improve readability [chapter 6, "Writing in Plain English," in Off. Invest. Educ. Assist. (1998)]. In their words, "No matter how sophisticated your audience is, if you use personal pronouns the clarity of your writing will dramatically improve" (Off. Invest. Educ. Assist. 1998, p. 22).[6] Examples of first person plural and second person singular personal pronouns are *we*, *us*, *our*, *ours*, *you*, *your*, and *yours*.

Following the publication of *A Plain English Handbook*, researchers often tabulate the count of first person plural and second person singular personal pronouns as a measure of writing clarity or readability. For example, the readability variable of Asay, Libby & Rennekamp (2018b) counts the number of personal pronouns using LIWC2015 software as one of its components. Higher counts of personal pronouns imply better readability. In an experimental setting, having personal pronouns in a business document that is only a few paragraphs in length might better engage the

[6]In an experimental setting, Asay, Libby & Rennekamp (2018a) identify a possible unintended consequence of having the SEC encourage personal pronoun usage in business disclosures. They find that a higher count of personal pronouns increases the reaction by retail investors to the financial disclosure.

reader in the material. However, should we expect improved readability with personal pronoun usage in a much larger annual report?

As a quick example, in the Form 10-K filed on February 24, 2017, by General Electric (GE 2017), 2.52% of all words are personal pronouns. The token *we* occurred 833 times. One sentence of GE's annual report actually used the word three different times, "With respect to manufacturing operations, *we* believe that, in general, *we* are one of the leading firms in most of the major industries in which *we* participate" (italics added) (GE 2017, p. 18). GE's report contained 76,272 words. Researchers have found little evidence to suggest that these stylistic changes made financial documents better at conveying valuation relevant data.

6.6. What Is Readability?

Ignoring our criticism of the mechanics associated with the application of readability measures in financial documents, we need to carefully consider exactly what we are attempting to capture. The Fog Index and other traditional measures of readability were designed and primarily used to create a grade-level measure for textbooks. In the context of financial documents, what is a desirable level of readability? Surely, making a 10-K accessible to someone in middle school is not the target. At the other end of the spectrum, financial jargon is shown in Loughran & McDonald (2014a) to be positively related to readability as measured by post–10-K stock volatility and analyst forecast errors. That is, the information from 10-K filings of firms using more complex and sophisticated financial terms seems to be better assimilated in prices and the forecasts of analysts. The broader literature on readability has always argued that readability must be designed in the context of the targeted audience (for example, see Davison & Kantor 1982). Should SEC filings be designed to be interpretable by professional analysts or average retail investors? These issues make the objective of document readability unclear.

Additionally, the use of pedantic words and writing style are not differentiated characteristics in any sample of financial filings. What appears as more complex writing in financial reporting is more often a reflection of the special vocabulary of an industry—such as pharmaceuticals, where chemical names are predominant—or the choice of some firms to include legal documents (leases, employee contracts, etc.) as part of the filing.

A more fundamental criticism of measuring readability is that like other measures of accounting quality, the document is simply a reflection of the firm and its structure (see Leuz & Wysocki 2016). Its composition may not be primarily determined by the strategic style of the author and instead may simply be a reflection of the characteristics of the firm.

All measures of readability, even those with some empirical or regulatory support, are focusing on an undifferentiated aspect of financial filings and may be measuring an attribute that is not important for the most relevant audience. More importantly, we argue in the next section that these measures are indirectly providing a noisy proxy for a firm attribute that, properly measured, would be useful in many empirical applications.

7. FIRM COMPLEXITY

Firm size is one of the most ubiquitous variables in empirical financial economics. Its inclusion as a control variable is intuitively obvious, but its actual specification (market capitalization or total assets) and measurement form (linear or log transform) are rarely explicitly dictated by a theoretical framework. Loughran & McDonald (2020) argue that firm complexity, broadly defined, is another important distinguishing attribute of companies that, like size, should be controlled for in the cross section. Typically it is not included as a control (or is proxied in a very specific context,

such as using the number of subsidiaries to predict audit fees), because it is a characteristic of the firm that cannot be precisely defined and is difficult to measure in its broadest sense.

Although one would expect complexity to be correlated with firm size, they argue that it is a distinct and differentiated aspect of firms. Although Netflix and McDonalds might be about the same market value, they are very different in terms of the complexity of their operating environment. Readability measures to some extent are simply a reflection of the underlying firm and likely are capturing some dimensions of a firm's complexity.

A firm's complexity is an artifact of many factors, for example, product heterogeneity, management hierarchy, acquisitiveness, or financial engineering. It is a recognizable attribute of the firm but is also broad and amorphous. Perhaps this is an instance where clear advantages exist for using textual methods to create an omnibus proxy for this attribute, where effective quantitative measures are not available.

Loughran & McDonald (2020) create a list of more than 300 words that are markers for firm complexity. Among the most frequently occurring tokens on their list are *subsidiaries*, *lease*, *acquisition*, and *foreign*. They then measure complexity using the number of unique occurrences of complex words in a firm's 10-K filing (i.e., the number of complex words occurring at least once). The litmus test for their proposed measure is based on predicting audit fees, where, from a vast array of prior research, firm size and complexity are primary determinants. In the context of audit fees, they show that their measure dominates other traditional control variables used in the audit fee literature. We would argue that this is a more relevant and differentiating feature in the cross section of firms than the readability of their financial documents.

8. CONCLUSION

We review some of the more recent contributions to the textual analysis literature in the broad field of financial economics, with an emphasis on papers that show its applicability in areas beyond those considered in previous studies. We also consider briefly the debate of humans versus machines in creating sentiment lexicons, arguing that the nuance of words makes humans the more effective arbiter of tone.

We then consider the topic of readability because its measure is controversial for mechanical and philosophical reasons. More importantly, we argue that careful consideration of what we attempt to measure with readability leads us to consider complexity as an important firm attribute. Historically, complexity has been measured in only a limited context, and yet it is an important and differentiating aspect of the firm.

Textual analysis has clearly become a relatively common arrow in the empirical quiver of financial researchers. In all research, the availability of empirical measures tends to frame and to some extent limit our thinking about the underlying concepts. Textual analysis, as its applications broaden, could give financial researchers ways of measuring relevant economic variables that historically have been difficult or impossible to capture using traditional quantitative data.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Andrew Lo (Co-Editor), Lynnette Purda, and an anonymous referee for helpful comments.

LITERATURE CITED

Agrawal S, Azar PD, Lo AW, Singh T. 2018. Momentum, mean-reversion, and social media: evidence from StockTwits and Twitter. *J. Portf. Manag.* 44:85–95

Asay HS, Libby R, Rennekamp K. 2018a. Do features that associate managers with a message magnify investors' reactions to narrative disclosures? *Account. Organ. Soc.* 68–69:1–14

Asay HS, Libby R, Rennekamp K. 2018b. Firm performance, reporting goals, and language choices in narrative disclosures. *J. Account. Econ.* 65:380–98

Azar PD, Lo AW. 2016. The wisdom of Twitter crowds: predicting stock market reactions to FOMC meetings via Twitter feeds. *J. Portf. Manag.* 42:123–34

Baker M, Wurgler J. 2006. Investor sentiment and the cross-section of stock returns. *J. Finance* 61:1645–80

Baker SR, Bloom N, Davis SJ. 2016. Measuring economic policy uncertainty. *Q. J. Econ.* 131:1593–636

Bartov E, Faurel L, Mohanram PS. 2017. Can Twitter help predict firm-level earnings and stock returns? *Account. Rev.* 93:25–57

Baxamusa M, Jalal A, Jha A. 2018. It pays to partner with a firm that writes annual reports well. *J. Bank. Finance* 92:13–34

Berelson BR. 1952. *Content Analysis in Communication Research*. Glencoe, IL: The Free Press

Blankespoor E, Miller GS, White HD. 2013. The role of dissemination in market liquidity: evidence from firms' use of Twitter. *Account. Rev.* 89:79–112

Bodnaruk A, Loughran T, McDonald B. 2015. Using 10-K text to gauge financial constraints. *J. Financ. Quant. Anal.* 50:623–46

Bonsall SB IV, Leone AJ, Miller BP, Rennekamp K. 2017. A plain English measure of financial reporting readability. *J. Account. Econ.* 63:329–57

Brown NC, Crowley RM, Elliott WB. 2020. What are you saying? Using *topic* to detect financial misreporting. *J. Account. Res.* 58(1):237–91

Burks JJ, Cuny C, Gerakos J, Granja J. 2018. Competition and voluntary disclosure: evidence from deregulation in the banking industry. *Rev. Account. Stud.* 23:1471–511

Burks L, Miller M, Zadeh R. 2014. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. In *Proceedings of the 10th National Conference on Earthquake Engineering*. Anchorage, AK: Front. Earthq. Eng.

Bushee BJ, Gow ID, Taylor DJ. 2018. Linguistic complexity in firm disclosures: obfuscation or information? *J. Account. Res.* 56:85–121

Bybee L, Kelly BT, Manela A, Xiu D. 2019. *The structure of economic news*. Work. Pap., Yale Univ., New Haven, CT

Chakrabarty B, Seetharaman A, Swanson Z, Wang X. 2018. Management risk incentives and the readability of corporate disclosures. *Financ. Manag.* 47:583–616

Chen CYH, Despres R, Guo L, Renault T. 2019. *What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble*. Work. Pap., Univ. Glasgow, Scotl.

Chen JV, Nagar V, Schoenfeld J. 2018. Manager-analyst conversations in earnings conference calls. *Rev. Account. Stud.* 23:1315–54

Chychyla R, Leone AJ, Minutti-Meza M. 2019. Complexity of financial reporting standards and accounting expertise. *J. Account. Econ.* 67:226–53

Cohen L, Malloy C, Nguyen Q. 2020. Lazy prices. *J. Finance* 75(3):1371–415

Das SR. 2014. *Text and Context: Language Analytics in Finance*. Found. Trends Finance Ser. Vol. 8, No. 3. Boston: Now Publ.

Das SR, Chen MY. 2007. Yahoo! for Amazon: sentiment extraction from small talk on the web. *Manag. Sci.* 53:1375–78

Davison A, Kantor R. 1982. On the failure of readability formulas to define readable texts: a case study from adaptations. *Read. Res. Q.* 17:187–209

Dechow PM, Ge W, Larson CR, Sloan RG. 2011. Predicting material accounting misstatements. *Contemp. Account. Res.* 28:17–82

Dyer T, Lang M, Stice-Lawrence L. 2017. The evolution of 10-K textual disclosure: evidence from Latent Dirichlet Allocation. *J. Account. Econ.* 64:221–45

Edmans A, Goncalves-Pinto L, Groen-Xu M, Wang Y. 2018. Strategic news releases in equity vesting months. *Rev. Financ. Stud.* 31:4099–141

Elliott WB, Grant SM, Hodge FD. 2018. Negative news and investor trust: the role of $Firm and #CEO Twitter use. *J. Account. Res.* 56:1483–519

Engelberg J, Henriksson M, Manela A, Williams J. 2019. *The partisanship of financial regulators*. Work. Pap., Univ. Calif., San Diego

Fama EF, French KR. 1993. Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* 33:3–56

Fama EF, French KR. 1997. Industry costs of equity. *J. Financ. Econ.* 43:153–93

Froot K, Kang N, Ozik G, Sadka R. 2017. What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *J. Financ. Econ.* 125:143–62

GE (Gen. Electr.). 2017. *Form 10-K*. Next-Generation EDGAR Syst., filed Feb. 24. U.S. Secur. Exch. Commis., Washington, DC. **https://www.sec.gov/Archives/edgar/data/40545/000004054517000010/0000040545-17-000010-index.htm**

Gentzkow M, Kelly B, Taddy M. 2019. Text as data. *J. Econ. Lit.* 57:535–74

Gentzkow M, Shapiro JM. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78:35–71

Gentzkow M, Shapiro JM, Taddy M. 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica* 87:1307–40

Griffin PA. 2003. Got information? Investor response to Form 10-K and Form 10-Q EDGAR filings. *Rev. Account. Stud.* 8:433–60

Guay W, Samuels D, Taylor D. 2016. Guiding through the fog: financial statement complexity and voluntary disclosure. *J. Account. Econ.* 62:234–69

Hayo B, Henseler K, Rapp MS. 2019. *Complexity of ECB communication and financial market trading*. Work. Pap., Univ. Marburg, Ger.

Hoberg G, Lewis C. 2017. Do fraudulent firms produce abnormal disclosure? *J. Corp. Finance* 43:58–85

Hoitash R, Hoitash U. 2018. Measuring accounting reporting complexity with XBRL. *Account. Rev.* 93:259–87

Hope OK, Wang J. 2018. Management deception, big-bath accounting, and information asymmetry: evidence from linguistic analysis. *Account. Organ. Soc.* 70:33–51

Hu M, Liu B. 2004. Mining and summarizing customer reviews. In *KDD-2004: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–77. Seattle, WA: Assoc. Comput. Mach.

Jiang F, Lee J, Martin X, Zhou G. 2019. Manager sentiment and stock returns. *J. Financ. Econ.* 132:126–49

Ke ZT, Kelly BT, Xiu D. 2019. *Predicting returns with text data*. Work. Pap., Becker Friedman Inst. Econ., Univ. Chicago

Kearney C, Liu S. 2014. Textual sentiment in finance: a survey of methods and models. *Int. Rev. Financ. Anal.* 33:171–85

Kim C, Wang K, Zhang L. 2019. Readability of 10-K reports and stock price crash risk. *Contemp. Account. Res.* 36:1184–216

Larcker DF, Zakolyukina AA. 2012. Detecting deceptive discussions in conference calls. *J. Account. Res.* 50:495–540

Leuz C, Wysocki P. 2016. The economics of disclosure and financial reporting regulation: evidence and suggestions for future research. *J. Account. Res.* 54:525–622

Li F. 2008. Annual report readability, current earnings, and earnings persistence. *J. Account. Econ.* 45:221–47

Li F. 2010. Survey of the literature. *J. Account. Lit.* 29:143–65

Lo K, Ramos F, Rogo R. 2017. Earnings management and annual report readability. *J. Account. Econ.* 63:1–25

Loughran T, McDonald B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66:35–65

Loughran T, McDonald B. 2014a. Measuring readability in financial disclosures. *J. Finance* 69:1643–71

Loughran T, McDonald B. 2014b. Regulation and financial disclosure: the impact of plain English. *J. Regul. Econ.* 45:94–113

Loughran T, McDonald B. 2016. Textual analysis in accounting and finance: a survey. *J. Account. Res.* 54:1187–230

Loughran T, McDonald B. 2017. The use of EDGAR filings by investors. *J. Behav. Finance* 18:231–48
Loughran T, McDonald B. 2020. *Measuring firm complexity*. Work. Pap., Univ. Notre Dame, Notre Dame, IN
Loughran T, McDonald B, Pragidis I. 2019. Assimilation of oil news into prices. *Int. Rev. Financ. Anal.* 63:105–18
Manela A, Moreira A. 2017. News implied volatility and disaster concerns. *J. Financ. Econ.* 123:137–62
Mayew WJ, Venkatachalam M. 2012. The power of voice: managerial affective states and future firm performance. *J. Finance* 67:1–43
Murphy PR, Purda L, Skillicorn D. 2018. Can fraudulent cues be transmitted by innocent participants? *J. Behav. Finance* 19:1–15
Off. Invest. Educ. Assist. 1998. *A Plain English Handbook: How to Create Clear SEC Disclosure Documents*. Washington, DC: U.S. Secur. Exch. Commis. **https://www.sec.gov/pdf/handbook.pdf**
Smales LA, Apergis N. 2017. Understanding the impact of monetary policy announcements: the importance of language and surprises. *J. Bank. Finance* 80:33–50
Soo CK. 2018. Quantifying sentiment with news media across local housing markets. *Rev. Financ. Stud.* 31:3689–719
Tetlock PC. 2007. Giving content to investor sentiment: the role of media in the stock market. *J. Finance* 62:1139–68

Annual Review of Financial Economics

Volume 12, 2020

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Financial Economics* articles may be found at http://www.annualreviews.org/errata/financial