Computational Analysis of Financial Narratives: Overview, Critique,
Resources and Future Directions

Mahmoud El-Haj*

Paul Rayson*

Martin Walker#

Steven Young†

Vasiliki Simaki‡

This draft: May 2018

Preliminary and incomplete: not for circulation outside the JBFA Capital Markets Conference
without author permission

† Corresponding author: Lancaster University Management School, Lancaster University, U.K., email:
s.young@lancaster.ac.uk * School of Computing and Communications, Lancaster University, U.K. # Alliance Manchester Business School, Manchester University, U.K. † Department of Linguistics and English Language, and Centre for Corpus Approaches in Social Science (CASS), Lancaster University. We are grateful for comments and guidance from Peter Pope and Andrew Stark (editors). We are also grateful to Paulo Alves, José Costa, Andrew Moore and Steve Whatham. Support for support with data and software resources described herein. Research support was provided by. Financial support was provided by the Economic and Social Research Council (ESRC) (contracts ES/J012394/1, ES/K002155/1, ES/R003904/1 and ES/S001778/1) and the Research Board of the Institute of Chartered Accountants in England and Wales.
Computational Analysis of Financial Narratives: Overview, Critique, Resources and Future Directions

Abstract

The objectives of the paper are two. We critically evaluate the strengths and weaknesses of research in accounting and finance applying methods from computational linguistics to study financial text. We conclude that while rigorous and transparent application of automated textual analysis methods provides significant opportunities, extant research risks overstating the magnitude of its incremental contribution (relative to manual analysis). Second, we discuss the applicability of a series of established methods from the natural language processing and corpus linguistics literatures that have yet to gain traction in an accounting and finance context but which we believe offer significant potential. As part of the discussion, we signpost relevant data and resources to support application of these techniques by researchers new to the area.
Computational Analysis of Financial Narratives: Overview, Critique, Resources and Future Directions

1. Introduction

Information is the lifeblood of markets and the amount of data available to decision-makers is increasing exponentially, with 90% of global information created during the last decade (Bank of England 2015), of which unstructured data (i.e., free-form text) is estimated to account for up to 80% of the content available on computer networks.\(^1\) The full information value of this resource cannot be exploited by governments, businesses, public services, and academics unless humans read these texts or devise an alternative means to derive information value from them. The dramatic growth in unstructured data is clearly evident in financial markets. For example, Dyer et al. (2017) report a 113% increase in the median length of U.S. registrants’ 10-K annual reports over the period 1996-2013, while El-Haj et al. (2018) report similar growth for the narrative component of U.K. firms’ annual reports published as PDF files. Unsurprisingly, accounting and finance researchers and increasing turning to solutions from computational linguistics to help process and analyze these data.\(^2\)

This paper has two objectives. First, we critically evaluate the strengths and weaknesses of research in accounting and finance applying methods from computational linguistics to study financial text. Second, we discuss the applicability of a series of established methods from the

---

\(^1\) Coen [https://www.quora.com/Why-is-natural-language-processing-considered-important](https://www.quora.com/Why-is-natural-language-processing-considered-important). In Merrill Lynch 1998 projected that available data will expand to 40 zettabytes by 2020 and estimated somewhere between 80-90% of all potentially usable business information may originate in unstructured form ([http://www.b-eye-network.com/view/6311](http://www.b-eye-network.com/view/6311)). Although the estimate includes video and image data as well as structured data in databases, the majority compromises plain text.

\(^2\) Throughout this paper we use the term “computation linguistics” to refer to the broad areas of natural language processing (NLP) and information retrieval from computer science, plus the smaller group of empirical methods developed in the field of corpus linguistics that combine manual and automatic approaches to the study of language.
natural language processing (NLP) and corpus linguistics literatures that have yet to gain traction in an accounting and finance context but which we believe offer potential. Third, we present two case study applications of financial textual analysis designed to promote future research by documenting the key steps involved in the analysis and generating two novel datasets.

Several studies already review aspects of the textual analysis literature in accounting and finance. Li (2010) discusses the benefits of computational content analysis over manual scoring and reviews the first wave of studies that use automated methods to examine accounting disclosures. Loughran and McDonald (2016) extend Li (2010) by combining an updated review of the mushrooming literature with a more focused survey and description of topics and methods that characterize extant work in accounting and finance. In particular they critique work on readability, highlight the importance of transparency when describing the process of converting raw text to quantitative measures, and reiterate Li’s (2010) call for economic theory to drive choice of textual analysis methods rather than vice versa. Kearney and Liu (2014) narrow the focus further by reviewing work on sentiment analysis published in finance before 2013.

Prior surveys start from the premise that the case supporting computational analysis of financial text is compelling. This view is not universally accepted, however, although the basis for scepticism is not clearly articulated. We identify and critique four views that may explain this suspicion: doubt over the value of studying narrative disclosures; distrust of computational approaches to scoring text; cynicism about the validity of applying computational methods to financial market disclosures; and concern over how the methods are applied and the relevance of the research questions examined. We conclude that the final explanation represents the most credible argument against automated analysis of financial textual analysis. We therefore proceed to evaluate extant research in light of this concern using three distinct but complementary lenses.
Our first evaluation lens compares the application of methods in accounting and finance research to four core principles and practices that underpin the computational linguistics approach. Our approach differs from prior surveys that define the textual analysis landscape according to the state-of-the-art in the accounting and finance literature. Armed with this general analytical framework and supplemented by additional information on specific textual analysis methods, we assess the extent to which the accounting and finance literature complies with standard operating procedures in computational linguistics. We conclude that while beacons of best practice can be found in accounting and finance research, poor practice is commonplace.

Our second evaluation lens is the suite of advantages to computational linguistic analysis (over manual coding) proposed by Li (2010). The predicted benefits include lower costs of scoring textual content, greater generalizability, higher objectivity, improved replicability, enhanced statistical power, and the ability to identify “hidden” linguistic features. Our review of prior research leads us to conclude that these benefits are less prominent than proponents of automated methods suggest. First, computational linguistics methods do not necessarily help to economize on the degree of manual coding. Applying these methods also involves significant researcher subjectivity that often lacks transparency, thereby constraining replicability benefits. A desire to minimize text retrieval costs also restricts generalizability to the extent that the literature focuses on a narrow set of text sources that are relatively easy to process (i.e., 10-K annual reports and earnings announcements published by U.S. registrations, and conference call transcripts). Further, the benefits of enhanced statistical power are also questionable given the degree of measurement error associated with naïve bag-of-words approaches that fail to reflect context and meaning. Finally, given the literature’s predominant focus on measures derived from simple word frequencies, only a small fraction of studies capture latent linguistic features.
Our third evaluation lens focuses on the magnitude of the incremental contribution offered by automated textual analysis of financial narratives over prior research employing manual content analysis. We document examples where use of computational methods provides significant new insights. However, we also show that conclusions offered by some of the so-called pioneering studies using automated scoring methods risk overstating their contribution insofar as they report evidence consistent with conclusions presented in earlier work relying on manual scoring methods. Further, we show that extant research using manual coding methods offers significant incremental insights that are unlikely to be superseded by studies using automated methods given the complex nature of the research questions and the corresponding demand for deep semantic parsing.

Collectively, our review suggests that while rigorous and transparent application of computational methods provides significant opportunities in the field of financial text, extant research risks overstating the magnitude of its incremental contribution relative to manual analysis. A key conclusion emerging from our analysis is that computational methods and high quality manual content analysis represent complementary approaches to studying financial text.

In addition to taking a more critical and dispassionate approach to evaluating the contribution of automated textual analysis research in accounting and finance, we extend Li (2010) and Loughran and McDonald (2016) by adopting a forward-looking perspective on textual analysis methods and their applicability. Specifically, rather than focusing solely on the limited set of methods employed to date in literature, we review established methods from the computational linguistics field that are yet to gain traction among accounting and finance researchers but which offer interesting potential. These methods include parts-of-speech and semantic tagging, keyness, named entity recognition, summarization, and text classification.
techniques. As well as helping to extend research horizons in textual analysis, the review speaks to the question posed by Loughran and McDonald (2016: 1223) regarding the potential benefits of parsing more deeply for contextual meaning in a business context. Their concern is that using more complex methods beyond simple word counts that ignore the sequence in which words are presented (i.e., meaning) may add more noise than signal to the empirical construct. We discuss tools from computational linguistics specifically designed to improve the signal-to-noise ratio by disambiguating word meaning, together with methods that permit researchers to address important questions that cannot be examined easily using a simple word-level strategy.

2. Analysing narrative content in accounting and finance research

Narrative disclosures have long attracted the interest of researchers in accounting and finance. However, the need to hand collect and manually score narrative content has constrained work in this area. European researchers and journals traditionally led the way in the field of manual content analysis (see Jones and Shoemaker 1994, Merkl-Davies and Brennan 2007, Beattie 2014), although a significant body of work also relies on manual methods to study aspects of U.S. narratives (e.g., Lang and Lundholme 1996 and 2000, Botosan 1997, Francis et al. 2002, Asquith et al. 2005, Sun 2010). Merkl-Davies and Brennan (2007) and Beattie (2014) provide comprehensive surveys of the accounting disclosure literature prior to the widespread adoption of automated textual analysis methods.

Early work by Abrahamson and Amir (1996), Das and Chen (2004), Henry (2006, 2008), Tetlock (2007), Tetlock et al. (2008), Li (2008) and Loughran and McDonald (2011) using automated methods to classify financial disclosures paved the way for the recent upsurge in papers applying computational linguistics methods to study the properties and economic
consequences of qualitative information. Proponents view these methods as offering a cost effective solution to the perceived weaknesses of manual content analysis including high subjectivity and low replicability, limited generalizability, and low statistical power (Li 2010). The adoption of automated scoring techniques by mainstream accounting and finance research is helping to move the literature away from a fixation on quantitative data towards a more balanced treatment of different information categories. However, many in the field remain sceptical about the merits of studying financial narratives through the lens of computational linguistics, although the basis for these concerns has not been clearly articulated in the prior literature. Understanding these concerns is the first step to objectively critiquing the contribution of research employing automated textual analysis methods.

We propose four non-mutually exclusive reasons for scepticism about the benefits of applying computational linguistics methods to study qualitative financial disclosures. The first reason reflects doubts about the incremental information value of qualitative financial disclosures over quantitative data. Scepticism may also reflect general distrust of claims that automated methods are able to provide any meaningful insights about such a complex and subtle phenomenon as language. In other words, any reliable analysis of text must involve manual reading and coding by default. The third reason to be suspicious of the computational linguistics approach is that domain-specific feature of financial text render it less amenable to automated processing than other corpora. Finally, concern over the way automated methods are applied and the (ir)relevance of the research questions examined may drive scepticism. We examine each of these four positions in further detail below.

2.1 Doubt over the value of qualitative disclosures
Distrust of applying computational linguistics approaches to study financial text may reflect deeper concern about the incremental value of qualitative information over quantitative data in financial market contexts. Summary quantitative measures such as asset prices, accounting earnings, and analyst forecasts impound information from multiple sources including qualitative disclosures. Further, quantitative data are often perceived as being more objective and verifiable. Collectively, these features could consign qualitative information to second order status in financial markets.

Scepticism over the relevance of qualitative disclosures in financial markets is nevertheless hard to reconcile with extant theoretical and empirical evidence. Theory establishes a clear link from mandatory and voluntary disclosure to efficient pricing and contracting (Beyer et al. 2010). All else equal, precision determines the value of information in a pricing context (Verrecchia 1983) and there is no reason why textual information must be systematically less precise than quantitative information. Empirically, a large literature documents the incremental usefulness of qualitative financial market disclosures in relation to: explaining contemporaneous stock returns (Francis et al. 2002, Asquith et al. 2005), predicting future earnings (Kothari et al. 2009, Feldman et al. 2010) and assessing earnings quality (Abrahamson and Amir 1997, Li 2008 and 2010, Frankel et al. 2016); predicting GAAP violations (Goel et al. 2010, Larcker and Zakolyukina 2012), fraud (Purda and Skillicorn 2015) and bankruptcy (Smith and Taffler 2000); predicting short-term stock returns (Tetlock 2007, Chen et al. 2014); reducing information asymmetry (Lang and Stice-Lawrence 2015); and reducing the cost of capital (Botosan 1997). Further, research confirms that the information contained in qualitative disclosures is economically important as well as statistically significant. Overall, extant research provides incontrovertible support for the economic relevance of qualitative financial data. Accordingly,
doubts about the incremental importance of financial text does not provide a valid basis for dismissing the potential contribution of applying computational linguistics methods to study qualitative disclosures in a financial market context.

2.2 Concern about the validity of processing text automatically

Written text and spoken language represent highly sophisticated forms of communication where context is critical to a complete understanding of meaning. Further, even in situations where context is clearly defined and understood by both message sender and receiver, ambiguity in meaning is still commonplace. Given the complexities and subtleties involved in processing textual communication, cynicism over the ability of algorithms to extract information and meaning reliably is perhaps not surprising when the human brain often struggles to interpret such signals correctly.

Despite the undeniably challenging nature of the problem, research in computer science, natural language processing, and linguistics has consistently demonstrated from the 1950s onwards that it is possible to use algorithms and statistical procedures to measure the properties of text and extract information automatically. While pioneering studies in the field of computational linguistics focused on machine translation, advances in computing from the early 1970s onwards led researchers to tackle increasingly sophisticated problems based on artificial intelligence (Sparck Jones 2001). Today, research activity in the combined fields of natural language processing and corpus linguistics ubiquitous and underpins everyday activities such as surfing the web, using product reviews when shopping, and electing politicians. For researchers in accounting and finance to dismiss this entire body of work as either economically irrelevant or
academically misguided is not conceivable. We therefore disregard this explanation for scepticism over the application of computational methods to analysing financial text.

2.3 Concern that financial text is not amenable to automated processing

A large fraction of work in the field of computational linguistics focuses on nonspecialist corpora constructed from generic sources such as newspaper articles, product reviews, political speeches, social media posts and popular fiction. Generic, non-technical language that is free of jargon and idiosyncratic content is more amenable to automated processing by off-the-shelf tools, and little or no domain expertise is required to facilitate extraction, feature selection and evaluation. In contrast, financial text is complex, jargon-heavy and context-specific. For example, the Fog index of the average 10-K annual report exceeds 19.0, which classifies the document as unreadable according to the standard interpretation of the index (Li 2008). Consistent with this view, financial market regulators routinely highlight problems caused by complex financial disclosures and stress the need to simplify language (Securities and Exchange Commission 1998, Financial Reporting Council 2015).

The complex, domain-specific nature of financial text raises concern about applying standard text-scoring methods from computational linguistics. For example, Li (2010) discusses the classification challenges posed by the lack of business-specific dictionaries when attempting score to annual report tone. Loughran and McDonald (2011) demonstrate empirically that the widely used Harvard-IV-4 TagNeg (H4N) dictionary has low classification accuracy for annual report disclosures because almost three-quarters of the H4N list is typically not considered negative in financial contexts. More generally, Loughran and McDonald (2016: 1223) pose the question of whether advanced NLP procedures are applicable for studying business text.
While financial disclosures undoubtedly feature high levels of complexity and domain-specific content, these characteristics alone do not provide sufficient grounds to reject the potential contribution offered by studying financial text using computational linguistics. Empirically, extant work confirms that good levels of classification accuracy are achievable in financial text applications when appropriate methods and reasonable adjustments are applied (Li 2010, Loughran and McDonald 2011, Huang et al. 2014, Dikolli et al. 2017). More generally, computational linguistics methods are widely applied in other domains featuring similar or higher levels of technical content, including medicine and healthcare (Olaronke and Olaleke 2015), education (Burstein et al. 2017), and biodiversity science (Thessen et al. 2012). Accordingly, we do not believe this view represents a valid reason for dismissing work that applies computational methods to study financial text.

2.4 Doubt over the way computational linguistics methods are applied and the relevance of the research questions studied

Rather than questioning the economic relevance of financial narratives or the validity of the scientific method of automated text processing (even in complex domains), this explanation for scepticism speaks to concerns about implementation. Specifically, are accounting and finance researchers seeking answers to appropriate questions when they use computation methods to classify financial text, and are researchers following best practice procedures when applying these methods? Unlike the previous three explanations considered above, this concern is harder to dismiss based purely on logical argument and contradictory evidence from the computational linguistics field. Indeed, Li (2010) and Loughran and McDonald (2016) both hint at such concerns in their respective reviews. For example, both studies argue that future research needs
to focus more on economic fundamentals rather than seeking applications off-the-shelf textual methods from computational linguistics. Further, Loughran and McDonald (2016) emphasize the importance of exposition and transparency in the next phase of research in this area. While criticism of extant work is implicit in these calls to action, neither review critiques prior research directly on these dimensions. Accordingly, the validity or otherwise of concerns over the way textual analysis methods have been implemented in extant accounting and finance research remains an open question that warrants careful scrutiny.

We proceed in sections 3-5 to evaluate prior research in light of this concern using three distinct but complementary lenses. The first evaluation lens presented in the next section compares the application of methods in accounting and finance research to discipline norms in the computational linguistics literature. The second evaluation lens discussed in section 4 is the suite of advantages to computational linguistic analysis (over manual coding) highlighted by Li (2010). The third evaluation lens discussed in section 5 assesses the magnitude of the incremental contribution offered by automated textual analysis of financial narratives over prior research employing manual content analysis.

3. Assessing research against core principles in computational linguistics

This section assesses the body of textual analysis research in accounting and finance against four principles of computational linguistics that collectively help to determine good practice, replicability and application of the scientific method. The four principles on which we focus are corpus collection, annotation, machine learning, and evaluation. In each case we provide a brief summary of the broad principle followed by an assessment of how well the accounting and finance literature applies the principle.
3.1 Corpus creation

Arguably the most fundamental principle in computational linguistics is that research should be based on empirical evidence derived from real language examples. Computational linguists therefore invest significant effort constructing large corpora to support their experiments. The aim is to create corpora that are representative of a language or variety of language, and then release the data as open access for other researchers to use as a common basis for further investigation. In addition to building general corpora that are representative of the variety of styles and sources for a particular language (e.g. spoken and written, fiction and non-fiction, newspapers, books, letters, etc.), researchers also create corpora to study specific research questions such as use of metaphors in palliative care (Semino et al. 2018) and construction of professional identities in corporate mission statements (Koller 2011).

Corpus construction as practiced in computational linguistics is not an activity that mainstream empirical accounting and finance research has pursued to any significant degree. First, since individual studies focus almost exclusively on one or several specific types of disclosure such as annual reports, earnings press releases, conference calls, media articles, etc., construction of corpora that are representative of financial market language in general is rare. A notable exception is Kothari et al. (2009) which analyses text from a variety of financial market sources including annual reports, analysts’ reports, and the financial press. Even in this case, however, the approach does not follow standard corpus construction rules and even more

---

3 The Brown corpus (Francis and Kucera 1979) consisting of 500 samples, each of 2,000 words, from 1960s American written English was one of the first machine-readable text corpora. The Lancaster-Oslo/Bergen (LOB) corpus (Johansson et al. 1978) was compiled using the same sampling frame, again totalling one million words, to represent British written English of the 1960s. The British National Corpus (Burnard 2007) comprises 100 million words from the 1990s.
critically, the corpus is not released publicly for other researchers to study. The same holds for
the majority of extant work that creates corpora from a single document type (e.g., annual
reports) to study specific research questions. The only study of which we are aware that makes
raw text available to other researchers for further analysis is El-Haj et al. (2018) which constructs
ten non-mutually exclusive corpora for U.K. annual reports including the entire report narrative,
the management commentary section, the letter to shareholders, the governance statement, and
the remuneration report.

The absence of financial market corpora restricts linguistic endeavour in this area by
limiting replicability and incremental analysis. It also leaves open the question of whether using
wordlists developed for one document type (e.g., 10-K filings) to measure the same linguistic
feature in an alternative corpus type (e.g., earnings announcements, conference calls, etc.) is
valid. Davies and Tama-Sweet (2012) and Dikolli et al. (2014) report systematic differences in
linguistic features across different corpora, which calls into question the practice of applying
static wordlists to different documents. While evidence reveals that results may be sensitive to
choice between domain-specific wordlists and general language dictionaries (Loughran and
McDonald 2011, Henry and Leone 2016), no work to date examines whether this effect extends
to different document types within the financial market domain.

Corpus creation is also limited by the extant focus on U.S. financial market text.
Accordingly, even if researchers were to make their corpora available publicly with immediate

---

4 Copyright law may represent an impediment to corpus building. However, where a corpus includes samples of text
from an original document rather than the entire contents of the document then the ‘fair dealing’ exception to
copyright law will apply and the corpus can be made public. Further, text collected from documents provided under
license by a third party commercial vendor be covered by this rule as long as dissemination states clearly states that
the corpus can only be used for non-commercial research purposes.
effect, the resulting resources are not guaranteed to provide a representative measure of international financial language because financial market regulations and business practices vary across countries. These differences likely cause international variation in the linguistic properties of financial text even where the underlying language (e.g., English) is held constant. Consistent with this view, Brochet et al. (2016) demonstrate how the linguistic properties of non-U.S. firms’ English-language conference calls differ from standard business language spoken by native English speakers. It is possible that these spoken differences extend to written text such as annual reports and earnings announcements, in which case the validity of applying wordlists developed on U.S. corpora (e.g., Henry 2006, Loughran and McDonald 2011, Campbell et al. 2013) to study similar linguistic features in non-U.S. settings may be questionable.

Although a significant body of research applying corpus linguistics methods to study financial market disclosures has evolved in parallel to mainstream textual analysis in accounting and finance research, it is rarely cited within the domain. Work includes analysis of annual reports (Charteris-Black and Ennis 2001, Rutterford 2005, Wang et al. 2012), CEO letters (Conaway and Wardrope 2010, Dragsted 2014), hedge fund manager commentary (Bruce 2014), conference calls (Rogers 2000, Crawford 2010a and 2010b), and audit reports (Tian and Liang 2011). We discuss this work more extensively in section 5: for now we merely note that while corpus construction is a feature of research on financial market text conducted by computational linguists, this practice is yet to gain traction in mainstream accounting and finance research.

3.2 Annotation

Language description (in Linguistics) and automated language processing (in Computer Science and Artificial Intelligence) typically involve adding information to the original corpus.
This process is known as annotation or tagging (Garside et al. 1997) and can involve manual or automated procedures. Manual annotation is the starting point for many applications. Manual analysis is used to identify particular features of interest such as tone and attribution in financial text (El-Haj et al. 2016) and helps develop deeper understanding of the feature(s) of interest. The process often proceeds in an iterative fashion where initial features are refined further as understanding of the phenomenon improves. Once a feature is clearly defined and a robust manual measurement system exists, NLP tools using probabilistic or rule-based methods can be constructed to replicate the annotation automatically to an acceptable level of accuracy for a much larger corpus than could have been manually annotated in any reasonable amount of time.

Automated annotation involves labelling elements of text using a predetermined algorithm. A common type of automated annotation is part-of-speech (POS) tagging which classifies each word according to its major word class (e.g. noun or verb), along with additional information (e.g. singular or plural, common or proper noun). These labels are subsequently used as inputs to disambiguation and feature identification strategies. Other types of annotation label increasingly complex levels of linguistic information such as morphology, grammar, syntax, semantics, pragmatics, and discourse. However, whereas POS categories are generally well understood and their corresponding tagsets are stable, consensus over the set of appropriate labels declines and the accuracy of automatic approaches decreases for higher annotation levels.

---

5 Automated annotation procedures are derived from prior manual tagging work designed to develop word-tag dictionaries and probabilistic models of tag frequencies and sequences to be generated, which form inputs to models POS tagging accuracy to a level equivalent to human accuracy (Leech et al., 1994).
6 For example, in tone or sentiment analysis words can be labelled as positive, negative or neutral, or a more fine grained categorisation from theories of emotion can be employed. When annotating word meanings, there is no widely agreed set of tags or standard dictionary on which an NLP system should be based.
Annotation is inherently an interpretative and potentially subjective process. The best results are obtained when multiple human coders annotate the same small set of text independently, compare results, create or update annotation guidelines and resolve disagreements, and then continue to independently annotate a much larger set of texts. Such a process more thoroughly tests the suitability of the set of categories being annotated and the degree of consensus regarding these categories. Metrics such as Krippendorff’s alpha and Cohen’s kappa measure agreement between annotators. [See Artstein and Poesio (2008) for a survey of the relevant literature]. Low levels of agreement may indicate that the task is poorly defined or that case-law guidelines require further refinement (Leech and Smith 2000).

Annotation practices in accounting and finance applications of textual analysis vary widely. Manual annotation is the dominant method employed in the literature, possibly reflecting the high levels of domain-expertise required in many situations. Examples of studies using manual analysis to identify particular linguistic features include Li (2010) and Huang et al. (2014) for tone, Kravet and Muslu (2013) for risk disclosures, and Dikolli et al. (2017) for CEO integrity. More frequently, however, researchers rely on third party dictionaries. Moreover, automated annotation methods such as POS and semantic tagging do not feature in extant work. Failure to apply this fundamental NLP principle casts a shadow over the accounting and finance literature, particularly given the dominance of dictionary-based strategies to feature detection and the attendant disambiguation problems plaguing context-independent approaches.

Comparing manual annotation strategies in accounting and finance with best practice NLP guidelines reveals several vulnerabilities for the literature. First, we are unaware of any published study where multiple human coders identify features by annotating the same set of text independently (with measures of annotator agreement reported to determine replicability and the
extent to which the coding task is clearly defined). Second, studies do not publish detailed case-law guidelines that define the annotation task and ensure replicability. Instead (and at best), the preferred approach involves providing selective examples of text that illustrate application of the classification process. Collectively, this lack of transparency is inconsistent with the inherently interpretative and subjective nature of the annotation process. The degree of opacity also contrasts with studies employing manual content analysis methods, which typically assess intercoder concordance and report comprehensive details of coding rules to help alleviate concerns over subjectivity (e.g., Milne and Adler 1999, Breton and Taffler 2001, Salzedo et al. 2018).

3.3 Machine Learning

Formally, a computer program is said to learn from experience ($E$) with respect to some class of tasks ($T$) and performance measure ($P$) if its performance at tasks in $T$, as measured by $P$, improves with $E$ (Mitchell 1997). Machine learning algorithms learn from examples (training data) to find relationships, develop understanding, make decisions and evaluate confidence from data provided. The quality and quantity of training data is a critical determinant of the success of the learning process. The goal of the training phase is to build a model capable of working with new data to a sufficient level of accuracy. Determining the appropriate level of classification accuracy is a particular challenge with machine learning because performance is context specific. For example, while beating chance may be deemed satisfactory for fraud detection, a 98% accuracy rate for an aviation system may be considered unacceptable.

---

7 Machine learning is used in many common tasks and applications including fraud detection algorithms on PayPal, travel optimization tools on Google Maps and Uber Taxi, and recommendation systems on Netflix and Spotify.
Text classification is a common application of machine learning systems. Text classification assigns one or more classes to a unit of text (e.g. document, sentence or word level) according to its content. Examples include identifying sentiment (tone) and detecting themes (topics). Machine learning algorithms work by extracting features that are common between data points. A feature may be defined as an individual measurable property (e.g., word class, word length, number of syllables) or the specific characteristic of a phenomenon being observed (e.g., future tense keywords associated with forward-looking commentary). Since the quality of extracted features is the primary factor determining performance of the model, it is not always clear ex ante which particular learning approach will yield the best classification results.

Machine learning methods can be grouped into three broad categories: supervised systems, unsupervised systems, and semi-supervised systems. Supervised systems require human intervention and involve methods such as linear regression, logistic regression, Naïve Bayes, Support Vector Machines (SVM), Support Vector Regressor (SRV), Random Forest, K-Nearest Neighbors (KNN), and Decision Trees.\(^8\) Intervention can take various forms including labelled data for a sample of manually-coded positive and negative sentences (e.g., Li 2010) or selecting corpus features using criteria such as Information Gain (e.g., Goel et al. 2010). The main cost of developing supervised systems is the intervention time required. In contrast, unsupervised systems rely on algorithms that learn how to group unannotated data automatically without requiring human intervention. The key process involves grouping similar data points based on pattern matching or clustering (e.g., Balakrishnan et al. 2010, Frankel et al. 2016). Finally, semi-supervised systems involve a combination of both labelled and unlabelled data.

\(^8\) SMV and SVR are similar algorithms designed to address different problems. SVM is a classifier that predicts discrete categorical labels while SVR performs regression to predict continuous ordered variables.
Machine learning applications in accounting and finance are scarce and focus mainly on simple text classification problems. Examples include Li (2010) measuring tone in forward-looking MD&A sentences, Huang et al. (2014) classifying tone in analysts’ reports, Sprenger et al. (2014) measuring sentiment in stock-related twitter messages, and Buehimaier and Whited (2014) measuring financing constraints using MD&A content. All three approaches employ a Naïve Bayes classifier trained on samples constructed using keywords (Buehimaier and Whited 2014) or manual labelling (Li 2010, Huang et al. 2014). Goel et al. (2010) use a range of classifiers to predict fraudulent financial reporting using 10-K filings, with text features selected from the entire corpus using the Information Gain criterion to reduce dimensionality. One of the few examples in the accounting literature applying machine learning for prediction is Frankel et al. (2016) who use SVR to assess the explanatory power of MD&A and conference call narratives for accruals. [See Holberg (2016) for a critique.]

With the exception of Goel et al. (2010), a notable gap exists between accounting and finance classification applications of supervised machine learning and NLP best practice. First, Naïve Bayes is presented as the default (and sole) classifier choice. While using Naïve Bayes is justified on the grounds that the method is simple to use and exhibits consistently good classification accuracy even in applications where the conditional independence assumption is violated, evidence reveals that alternative algorithms may perform better (Domingos and Pazzani 1997). Theory provides no clear guidance about which classifier is likely to perform best in a given situation and therefore NLP studies typically examine a range of algorithms before

---

selecting the best performer. For example, Goel et al. (2010) use Naïve Bayes as a preprocessor to explore useful feature sets for SVM. Their results reveal that SVM displays superior classification accuracy and that their Naïve Bayes classifier fails to beat the random baseline.

Second, studies applying Naïve Bayes often provide incomplete implementation details. For example, two distinct but interconnected variations of Naïve Bayes are the multinomial model and multivariate Bernoulli model (McCallum and Nigam 1998). Ren et al. (2013) report differences in classification accuracy for the two models conditional on the size of the feature set, with the multinomial model over-fitting (performing better) when the feature set is small (large). While Goel et al. (2010) justify their decision to use the multinominal model, other studies are silent on model choice and associated rationale. A related implementation issue is the specific algorithm used to estimate the classifier. Information on the estimation algorithm is important for replication purposes. Goel et al. (2010: 34) use the Rainbow (Bow) classifier toolkit (McCallum 1996) while Li (2010: 1062) relies on the Algorithm::NaiveBayes module in Perl. Other studies, in contrast, do not discuss the specific algorithm used.

Feature selection is the key determinant of classification performance in machine learning applications. NLP applications often rely on a data driven approach to feature selection whereby an algorithm such as Information Gain is used to identify words and n-grams with higher discriminator or predictive power. Examples include Goel et al. (2010), Balakrishnan et al. (2010), and Frankel et al. (2016). Alternative approaches to constructing a training sample for classification rely on manual labelling (Li 2010, Huang 2014) and keywords (Buehimaier and Whited 2014). The advantage of these latter methods is that they incorporate domain expertise; the disadvantage is they may overlook other important features of the text that cannot be determined by the researcher ex ante. It remains an open question which approach provides
superior performance, whether a combined strategy offers incremental classification gains, and whether the best performing machine learning model can outperform methods that are easier to implement such as word-frequency measures (Henry and Leone 2016).

Finally, we caution against the practice of using a model trained on a specific corpus for a particular task to undertake a different classification task on content from a different text source. For example, Henry and Leone (2016) compare the accuracy of a Naïve Bayes classifier against a simple dictionary-based strategy for detecting tone in earnings announcements. Rather than build a classifier from scratch, Henry and Leone (2016) use the model developed by Li (2010) for classifying the tone in forward-looking MD&A sentences. Results demonstrate no incremental classification accuracy for the machine learning approach leading to the conclusion that “although more complex techniques are potentially advantageous in certain contexts… word-frequency tone measures are generally just as powerful in the context of financial disclosure and capital markets”, Henry and Leone (2016: 153). While such a conclusion may be valid, we note that the NLP literature raises doubts about the portability of machine learning models across corpora and classification tasks. Concerns over portability coupled with the narrow perspective on model selection adopted by Henry and Leone (2016) mean that it is premature to dismiss the value of automated classifiers in an accounting and finance context.

3.4 Evaluation

A crucial component of the computational linguistics endeavour is evaluation (Resnik and Lin 2013). Evaluation is required for all elements of the automated pipeline including corpus
extraction, corpus annotation and text classification.\textsuperscript{10} The task usually involves creating a test dataset that is manually coded with the expected results or annotation and represents the “gold standard” against which automated results are assessed. Metrics used to assess the performance of automatic annotation systems include simple accuracy (defined as the percentage agreement between the automatically assigned labels and the manually annotated gold standard) and error rate (defined as 100 minus the accuracy score). Where multiple outcomes are possible performance is typically assessed in terms of precision and recall (Junker et al. 1996, Manning and Schütze 2008). Precision measures the fraction of false positives (Type I errors) and is viewed as a measure of exactness or quality, while recall measures the fraction of false negatives (Type II errors) and reflects a measure of completeness or quantity. A measure of overall accuracy is also often computed using the F score, which is equal to the harmonic mean of precision and recall (Van Rijsbergen 1979).\textsuperscript{11}

The quality and transparency of evaluation in the accounting and finance literature is mixed. Formal assessments of retrieval accuracy and associated error rates are rarely reported, thereby creating the appearance that extraction is a reliable process similar to retrieving financial market data from COMPUSTAT and CRSP when this is often not the case.\textsuperscript{12} For example, Li

\begin{scriptsize}
\begin{enumerate}
\item While different evaluation approaches are applicable in different settings, several core principles apply. These general principles were surveyed originally in the Expert Advisory Group on Language Engineering Standards (EAGLES 1996). Intrinsic evaluation involves understanding the performance of a specific processing tool. It can be distinguished from extrinsic evaluation which seeks to determine how well a specific processing tool is performing as part of an entire language technology system. For example, how much difference does a POS tagger make to a complete speech recognition system that allows a user to talk to their smart phone?
\item The F score is derived such that $F_{\beta}$ measures the effectiveness of retrieval with respect to an individual who attaches $\beta$ times as much importance to precision as recall. The $F_1$ score places equal weight on precision and recall, whereas the $F_2$ ($F_{0.5}$) score weights recall (precision) higher than precision (recall).
\item El Haj et al. (2018) is one of the few studies in accounting and finance literature that provide detailed information on retrieval and classification accuracy for text extracted from U.K. annual reports published as digital PDF files.
\end{enumerate}
\end{scriptsize}
(2010: 1060, footnote 6) conducts a random check of 200 filings and documents success rates between 85% and 90% for 10-Ks.

Evaluation of text-derived proxies constructed using wordlists are more common although still not ubiquitous. A relatively small fraction of studies follow NLP best practice and conduct manual evaluations (e.g., Dikolli et al. 2017, Athanasakou et al. 2018). At the opposite end of the spectrum, formal evaluations are replaced by ad hoc “sanity checks” (Bozanic et al. 2013, Buehimaier and Whited 2014). However, the dominant approach to assessing construct validity involves regressing the text-derived proxy on firm- and market-level characteristics predicted by economic theory to covary with the latent linguistic feature (Li et al. 2013, Dikolli et al. 2017, Henry and Leone 2016, El-Haj et al. 2018). While this indirect evaluation method is capable of shedding light on the properties of text-derived constructs, the approach is problematic for two reasons. First, researchers cannot rule out the possibility that estimated associations reflect correlated omitted variable bias. In this regard it is noteworthy that econometric standards normally applied to tests of economic hypotheses do not appear to extend to assessments of construct validity. Second, correlations provide an inherently weak evaluation framework since it is not possible to specify the true value of the regression coefficient and therefore researchers are unable to assess the sign and magnitude of measurement error with any precision. In contrast, manual evaluations measured against a gold standard provide precise estimates of noise and bias.

A small but growing number of studies in accounting and finance apply machine learning classifiers (Li 2010, Huang et al. 2014, Buehimaier and Whited 2014). A key evaluation principle for machine learning applications is that text which has been used to train a system should not be used for testing. Separating the dataset into two groups to create a training and
holdout sample represents a simple approach to ensuring independence. A more sophisticated bootstrapping-type approach is K-fold cross validation whereby a corpus is split into K equal-sized sections. The experiment is then repeated K times and on each iteration a different section (or fold) is used for testing while the remainder are used for training. A typical partition in each iteration involves using 90% of the data for training and the remaining 10% for testing. The process iterates on a number of runs (folds) with random partitioning in each fold. The number of folds applied is usually between five and 10. Classification accuracy may also be assessed by comparing out-of-sample manual classification with automated classification, with results presented in a “confusion matrix” that reports correct classifications on the main diagonal and errors on the off-diagonal elements. The advantage of this approach is it provides evidence on bias as well as overall classification accuracy.

Variable practice is again evident in the literature. For example, Li (2010) and Huang et al. (2014) evaluate the classification accuracy rate of their Naïve Bayes algorithm using a 10-fold cross validation, whereas Buehimaier and Whited (2014) do not report formal validation tests. Evaluation strategies also vary significantly within individual studies. For example, while Li (2010) evaluates the accuracy with which his Naïve Bayes classifier captures the tone of forward-looking MD&A sentences, no evaluations are reported for the arguably more fundamental task of identifying forward-looking performance sentences via a simple wordlist.

---

13 Evaluation is so integral to NLP that many areas and subfields organise their own annual evaluation competitions with standard training data provided to all participants and hidden test data only released near the time of the conference or event. Between the releases teams of researchers compete using different algorithms to produce the best systems. Examples of such events include the Text Retrieval Conference (TREC), Message Understanding Conference (MUC), the Sense Evaluation and Semantic Evaluation conferences (Senseval and SemEval), and the Conference and Labs of the Evaluation Forum (formerly known as Cross-Language Evaluation Forum, CLEF).
approach (described in appendix B of his paper). Finally, none of the machine learning applications to date evaluate bias when assessing classification performance.

4. **Assessing automated scoring methods against the advantages proposed by Li (2010)**

Li (2010) compares manual versus automated content analysis. The advantages associated with manual analysis include more precise coding and scope for more detailed and tailored analysis. However, two structural problems associated with manual analysis offset these advantages. First, the cost of collecting data tends to be very high, leading to small sample sizes that may limit the scope of empirical analysis in terms of generalizability and statistical power. Second, researcher subjectivity that is inevitably required in the manual coding process can limit replicability. Li (2010: 145) argues that automated approaches to analysing text help to resolve both problems and that as a result computer-based analysis of text offers significant incremental advantages over manual coding approaches. Specifically, once a scoring algorithm has been developed and validated, automated methods enable researchers to score textual content for extremely large samples of documents quickly at low marginal cost. All else equal, larger sample sizes increase statistical power and promote generalizability. Further, because it is possible to apply a scoring algorithm consistently across multiple documents, the approach eliminates conscious and unconscious researcher bias, while also helping to minimize random measurement error associated with inconsistent application of manual coding rules. The approach should also promote replication and extension where complete details of the scoring algorithm and the process used to implement the analysis are reported transparently.
At first sight, the advantages proposed by Li (2010) appear unequivocal and uncontroversial. However, closer inspection highlights the conditional nature of these benefits as discussed below.

4.1 Reduce data collection costs

Data collection costs are determined by the text preprocessing, classification, and construct validation requirements associated with the specific research question and the corresponding research design. Consistent with Li’s (2010) claim that automated methods facilitate analysis of very large samples of text at relatively lower cost, extraction costs are indeed low for studies employing documents collected from commercial platforms that offer a reliable application programming interface (API) or from a document repository such as EDGAR whose website architecture is designed to support web scraping. Similarly, preprocessing costs are relatively low for documents presented in plain text or HTML format with either no structure (e.g., social media posts or media articles) or a highly standardized reporting template (e.g., 10-K and 10-Q fillings). Scoring costs are also negligible when classifying text using simple word frequencies based on off-the-shelf keyword lists (e.g., measuring tone) and using well-established algorithms for measuring features such as document length, readability (e.g., fog, Flesch, and bog indices) and text reuse (e.g., cosine similarity). Finally, validation costs are limited when statistical correlations rather than manual procedures are used to evaluate classification performance (e.g., Li et al. 2013, Henry and Leone 2016).

However, automated text scoring procedures do not guarantee a material reduction in data collection costs because the conditions outlined above describe a limited set of very basic textual analysis applications. Relaxing these artificial constraints increases implementation costs
dramatically as discussed in section 3 in relation to annotation and evaluation, to the extent they could exceed those associated with a smaller sample manual coding design. As the following examples illustrate, processing challenging document formats such as PDF files, extracting specific sections from highly structured unstandardized documents, undertaking more sophisticated classification tasks involving significant domain expertise, and validating classification performance manually are high cost activities:

- Li (2010) uses supervised machine learning methods (naïve Bayes) to classify tone in forward-looking statements from the MD&A section of firms’ 10-K fillings. The training step involves the author and 15 research assistants coding 30,000 forward-looking sentences manually. Based on an optimistic estimate of 20 seconds to classify and record the average sentence, the training phrase of the process alone equates to 10,000 person-hours.

- El-Haj et al. (2018) develop and validate a tool to retrieve narrative content and document structure from UK annual reports published as PDF files, and then classify generic sections (e.g., Chair’s letter, management commentary, governance statement, remuneration report, financial statements) to facilitate time-series and cross-sectional comparisons. Their method involves locating the report table of contents, synchronizing page numbers in the native report with page numbers in the corresponding PDF, and then retrieving content separately for each section listed in the table of contents. The report table of contents is identified using a “gold standard” set of common section titles and associated synonyms identified using an initial sample of 50 reports selected at random and then refined manually via three iterations each based on training samples of 1,000 reports selected at random. These training samples are also used to (a) optimize the page synchronization algorithm by manually comparing page number allocation with report page numbers in the native PDF file and (b) develop and
refine separate keyword lists for classifying generic annual report sections. Finally, manual validations assessing retrieval and classification accuracy are performed on 11,009 separate sections from a random sample of 586 reports.

- Dikolli et al. (2017) construct a measure of CEO integrity by comparing explanations of firm performance in the letter to shareholders with those presented in the corresponding MD&A where language is more constrained by corporate lawyers. A version of the Linguistic Inquiry Word Count (LIWC) causation wordlist is used to identify “explanation sentences”. Three significant manual steps are involved in implementing the research design. The first step involves constructing the corpus of CEO shareholder letters manually from firms’ PDF annual reports because these disclosures are not available as an EDGAR filing. The starting point for the data collection procedure is the Compact Disclosure database which contains shareholder letters and MD&As in machine readable text formats. This involves traveling to the University of Pennsylvania’s Wharton Library where the Compact Disclosure monthly CD library is physically located. Since Compact Disclosure’s coverage declines after 2002, the sample is supplemented with shareholder letters retrieved manually from the Thomson ONE and OneSource databases, with text either copied directly from the PDF or converted to plain text using OmniPage 18 PDF conversion software on a document by document basis. The second manual step involves refining the original LIWC causation wordlist (which is not designed for analysis of business text) to increase the precision with which CEO explanations are identified (Loughran and McDonald 2011). The third manual step involves the authors validating the refined causation wordlist by reading a random sample of 5,000 shareholder letter sentences and demonstrating that the refined list predicts explanation sentences better than both chance and the original LIWC wordlist.
These three examples illustrate how automated textual analysis procedures do not necessarily economize significantly on manual collection costs. Only for the most trivial (and potentially least interesting) textual analysis applications is material manual input mitigated or avoided altogether. Typically, the level of manual intervention increases rapidly as the research question and the linguistic features of interest become more sophisticated. Ultimately, significant manual intervention by domain experts is an unavoidable feature of most disclosure research in accounting and finance irrespective of the specific methodology employed. The choice between manual versus automated coding approaches merely influences the stage in the research process where significant manual intervention occurs (along with the degree of coding precision and the scalability of the scoring process). Both approaches likely involve significant data collection and scoring costs when implemented rigorously, with the nature of the specific research question determining relative costs and benefits.

Although significant cost savings are unlikely to accrue to researchers who initially develop and validate new automated textual analysis resources in accounting and finance, substantial cost reductions are possible at the aggregate level where these resources are made available for other research teams to use and extend. Resources provided by Loughran and McDonald (2011 and 2016) and El-Haj et al. (2018) illustrate the scope for reducing data collection and processing costs offered by automated textual analysis methods. Provision of corpora, keyword lists, training data, and script for harvesting, processing and scoring text as part of an Open Science (European Commission 2018) approach to work in this area offers substantial benefits to the discipline and beyond. As research in this area matures, we hope that more sharing of text resources will become the norm rather than the exception.
While our discussion to date focuses on costs from the demand side, minimizing extraction and processing costs is (or at least should be) an increasingly important consideration for regulators charged with developing disclosure rules. Holding the intrinsic content of narrative reporting constant, the rapid growth in artificial intelligence and big data applications means the ability to automatically retrieve and process data on a large scale at low cost is emerging as a key factor influencing the usability of financially-relevant text. The form and structure of narrative disclosures therefore represents an additional consideration alongside the traditional focus on content for financial market regulators developing new disclosure rules. The recent International Accounting Standards Board (IASB) Discussion Paper on disclosure principles (IASB 2017) illustrates the scale of the challenge. The 107-page Discussion Paper contains no reference to the impact of technology on the way accounting disclosures will be used in the future. Rather, the proposals implicitly assume that manual reading by humans on a document-by-document basis will continue to be the primary method by which accounting information is accessed and processed. Unless regulators acknowledge the increasing importance of automated processing and take steps to reduce retrieval costs by improving accessibility, many of the benefits associated with higher quality disclosure may not be realized fully.

4.2 Increase statistical power via larger sample sizes

Even if computer-based approaches do not guarantee material cost savings, their scalability enables researchers to work with much larger samples than is feasible using manual methods. All else equal, larger samples increase statistical power. However, the positive link

\[14\] The U.S. experience with the EDGAR filing system coupled with standardized reporting templates such as form 10-K demonstrates how accessibility considerations drive information usage in a big data environment.
between statistical power and sample size assumes constant measurement error. The problem for many computer-based textual analysis applications is that measurement error increases dramatically relative to manual scoring due to the blunt nature of classification processes based on naïve features such as keywords, word length, and sentence length.15

Extraction errors are especially common when working with PDF files. For example, El-Haj et al. (2018) discuss how certain character strings in PDF files are systematically corrupted in the retrieval process. Failure to identify and correct these errors can induce significant measurement error in subsequent analyses based on word and syllable counts (e.g., readability), key word frequencies, and statistical feature extraction (e.g., Information Gain criterion or log likelihood ratios). Another common error when working with PDF files is text ordering. For example, content presented in column format can easily be scrambled if the extraction process incorrectly reads directly across the page rather than column by column. Such errors often depend on the native file type and the software used to generate the PDF file. Unfortunately, such errors are hard to detect and researchers must therefore be vigilant.16

Accordingly, gains in the signal-to-noise ratio from applying larger samples may be offset by greater measurement error due to noisier extraction and classification. Indeed, as the complexity of the text processing task increases it is perfectly possible that more precise manual coding applied to a few hundred observations may generate greater statistical power than

---

15 Dikolli et al. (2017) report error rates of 32% when classifying explanation sentences using a tailored wordlist. Li (2010) reports error rates up to 15% for extracting the MD&A section from the 10-K filing. Davis and Tama-Sweet (2012: footnote 14) note that 10-K and 10-Q filings extracted directly from EDGAR include a large number of extraneous characters due to different text-file formats used by registrants and EDGAR. Davis and Tama-Sweet use the 10-K Wizard tool to clean raw text files prior to analysis. Their approach contrasts with other studies including Li (2008) which use the raw files retrieved directly from EDGAR.

16 For example, where the unit of analysis is individual words then this type of error may have no impact on results because the order in which words are presented is unimportant. In contrast, the problem may create severe measurement error where sentences are the unit of analysis (including readability metrics).
automated methods applied to tens of thousands of documents. The idea that computer-aided analysis of very large samples of text mechanically increases statistical power is therefore a fallacy; and the allure of large samples associated with automated content analysis represents a potentially dangerous trap, particularly for less the experienced researcher.

4.3. Promote generalizability

The opportunity for researchers to work with much larger samples due to the scalability of automated textual analysis methods helps promote generalizability through analysis of more representative samples. While superficially such a claim seems uncontentious, restrictions on data extraction and text processing mean that even very large samples may still suffer from material selection bias. For example, research on 10-K and 10-Q filings typically omits firm-years where the MD&A is incorporated by reference to the PDF annual report. As a result, even very large samples of 10-K may not be fully representative of the population of filers where firms whose MD&A is incorporated by reference differ systematically on one or more dimensions. Similarly, El-Haj et al. (2018) analyse 19,526 UK annual reports published as digital PDF files from 2003 to 2016. Their inability to process image-based PDF files induces a potential time and size bias in the final sample since scanned PDFs were relatively common before 2006 and more frequently published by smaller firms.

The scope of potential generalizability benefits also depends critically on how one defines the population from which the sample is drawn. Where the population is defined as a specific reporting regime then a large sample that is more representative of that regime enhances

---

17 Dikolli et al. (2017) is a notable exception. They hand collect MD&As where this section is incorporated in the 10-K filing by reference to the PDF annual report.
generalizability. In contrast, where the population is defined to include multiple reporting regimes then simply increasing sample size from a single regime (e.g., the U.S.) does not necessarily produce insights that are more generalizable to the other (e.g., non-U.S.) regimes. When assessed against this broader notion of international generalizability, there is reason to believe that recent work employing automated textual analysis methods may actually produce results that are less generalizable than those produced using manual scoring methods. By way of illustration, consider recent work examining annual reports in general and management commentary in particular (e.g., Li 2008, Brown and Tucker 2011, Loughran and McDonald 2011, Lang and Stice-Lawrence 2015, Dyer et al. 2016). The ease with which 10-Ks filed by U.S. registrants can be harvested and processed means that the majority of extant research on annual report has naturally focused on such documents. Researchers’ ability to scrape and process the population of 10-K filings from EDGAR undoubtedly enhances generalizability within a U.S. context. However, since the 10-K reporting template produces annual reports that differ significantly in structure and content from those published by firms reporting in most other jurisdictions, the applicability of computer-aided analyses of 10-Ks to annual reports published outside the U.S. is unclear. Further, since the population of non-U.S. annual reports far exceeds the number of reports published using the 10-K format, it is questionable whether extant findings can be considered representative of the median annual report globally.

---

18 The same caveat may also hold the subset of U.S. registrants whose MD&A section of Form 10-K is incorporated by reference to their PDF annual report.

19 Samples sizes in studies examining 10-K annual reports vary dramatically. For example, Dyer et al. (2017) examine 75,991 reports over the period 1996-2013; Li (2008) examines 55,719 for the period 1994-2004; Loughran and McDonald (2011) example of sample of 37,287 reports for the period 1994-2008; and Brown and Tucker (2011) study a sample of 28,142 reports for the period 1997-2006. By comparison, the international sample of non-US annual reports drawn from 42 countries examined by Lang and Stice-Lawrence (2015) comprises 87,608 reports. Globally, therefore, the representative (median) annual report almost certainly does not have the same structure and content as an annual report filed on Form 10-K.
Because computational analysis of financial narratives relies on access to large volumes of text, the source, format and structure of text resources collectively influence retrieval costs, which in turn shape research agendas and ultimately generalizability. It is perhaps no accident therefore that extant research focuses on financial corpora where retrieval costs are low or moderate such as 10-Ks, earnings announcements and conference calls. While these sources represent natural text resources for accounting and finance researchers to exploit, high retrieval costs risk marginalizing other equally interesting and economically important types of financial narrative. For example, the following sources where retrieval costs are high have received much less attention in the literature: glossy (PDF) annual reports (Grüning 2011 and 2014, Lang and Stice-Lawrence 2015, El-Haj et al. 2018), analysts’ reports (Twedt and Rees 2012, Huang et al. 2014), initial public offering prospectuses (Ferris et al. 2013, Loughran and McDonald 2013), SEC litigation releases (Zaki and Theodoulidis 2013), announcements by politicians and financial market regulators (Wisniewski and Moro 2014); and peer-to-peer lending platforms (Gao and Lin 2015).

Generalizability may also be affected by the procedure used to extract text. For example, Lang and Stice-Lawrence (2015) conduct the first large scale analysis of PDF annual reports for an international sample comprising 87,600 documents for over 15,000 non-U.S. firms from 42 countries, Results reveal how higher quality disclosures proxied by IFRS adoption are associated with positive stock market outcomes. However, Lang and Stice-Lawrence (2015) do not distinguish between management commentary and financial statement footnotes because their text processing method does not capture the location of commentary within the report.20

---

20 The authors approach the task of analyzing highly structured PDF files by applying Xpdf and QPDF proprietary software to convert content to ASCII format and then using a pearl script to isolate running text.
Subsequent work by El-Haj et al. (2018) on U.K. annual reports separates management commentary from financial statement footnotes. Consistent with the narrow scope of IFRS, El-Haj (2018, Table 3) show that changes in narrative content following IFRS adoption are confined to the financial statements component of the annual report and that as a result conclusions in Lang and Stice-Lawrence (2015) do not generalize to management commentary.

The desire or need to minimize extraction costs risks introducing observational bias into the textual analysis literature in the form of a disproportionate focus on a subset of narrative communications. This bias is a manifestation of the well-known “streetlight effect” that occurs when individuals limit their search activity to locations where it is easiest to look. For example, the high concentration on 10-K and 10-Q filings in the literature is at odds with debate over the relevance and timeliness for economic decision making of annual reports (Chen and Li 2015).

Observational bias is also evident in the high concentration of research on U.S. firms’ narrative disclosures. Extant research provides consistent evidence that (quantitative) financial reporting outcomes are determined by preparers’ incentives and institutional features including legal origins, investor protection, governance arrangements, stock market development, enforcement arrangements, etc. (Hail et al. 2010). The same is almost certainly true for qualitative financial disclosures (Ernstberger and Gruning 2013, Lang and Stice-Lawrence 2015).

21 The effect is often illustrated using the principle of the drunkard’s search (Kaplan 1964): When a policeman observes a drunk man searching for something under a streetlight and asks what the drunk has lost, the drunk relies that he lost his keys and they both proceed searching together. After a few minutes the policeman asks if the drunk is sure he lost them here and the drunk replies that he actually lost them in the park. When quizzed by the policeman why he is searching under the streetlight, the drunk replies that the light is better here.

22 Evidence on the market reaction around the release of a U.S. company’s annual report and related 10-K filing is mixed. MD&A’s are considered deficient because they typically include boilerplate statements and immaterial detail but little information of substance (SEC 2003, Garmong 2007, Cole and Jones 2005). Further, while most research on 10-Ks focused on the MD&A, the SEC concludes that this component of the annual report is generally deficient because it is heavy on boilerplate statements and immaterial detail, and light on information of economic substance (SEC 2003, Garmong 2007, Cole and Jones 2005). See Stanton and Stanton (2002) for discussion of research on corporate annual reports.
Indeed, there is good reason to believe these differences may be more pronounced for narratives such as annual reports and earnings announcements because unlike financial statement outcomes, textual disclosures are not bound by the temporal discipline of double-entry bookkeeping. Accordingly, insights emerging from the rapidly growing textual analysis literature are likely to be partial at best.\textsuperscript{23}

4.4 Improve objectivity and replicability

Manual content analysis inevitably involves researcher subjectivity, which in turn raises questions about replicability. Studies applying manual coding methods typically address this challenge by providing extensive details of the scoring instrument together with examples that demonstrate how the coding procedure is applied. While these steps provide clarity and help to support replication, researchers are nevertheless required to apply significant judgement even when implementing the most transparent coding procedure. Automated content analysis procedures are considered attractive because algorithms apply a set of rules consistently and dispassionately. Results for a single document or an entire sample are therefore replicable multiple times and by different individuals.

While the promise of high replicability is an attractive feature of computer-based textual analysis approaches, closer inspection highlights the need for caution where this proposed

\textsuperscript{23} Progress on processing new document types is hampered by the need to develop new extraction methods. Unfortunately, accounting and finance researchers cannot rely on the computational linguistics literature to deliver required solutions because most NLP researchers do not possess the domain expertise required to parse and classify complex financial disclosures. The potential for a market failure therefore exists because pure method papers are rare in the accounting and finance literature generally and top-tier journals in particular. We believe such work offers significant upside benefits for journals because the resulting studies are likely to be cited widely both in accounting and finance, and in other disciplines (including computational linguistics) where demand exists for access to business-related text resources. Significant progress in this area almost certainly requires multidisciplinary teams and possibly cross-discipline coordination among journals.
advantage is concerned. First, automated scoring procedures are not immune to researcher subjectivity as indicated in section 3.3. Varying degrees of judgement are necessary in all but the most basic automated applications involving mechanical feature extraction (e.g., readability scores or word frequencies based on a pre-existing wordlist) applied to a corpus of text such as the MD&A component of firms’ 10-K filing. Examples of researcher subjectivity when applying computational linguistic approaches include: text preprocessing choices such as stemming and removal of stop words (Lourghan and McDonald 2016); construction of domain-specific wordlists (Loughan and McDonald 2011, Li et al. 2013, Dikolli et al. 2014, Athanasakou et al. 2018); manual coding of training data in machine learning applications (Li 2010, Huang et al. 2014); feature set selection using IG criterion (Goel et al. 2010); choice of estimation parameters in classification tasks using methods such as Naive Bayes, SVM, and topic modelling (Ball et al. 2015, Dyer et al. 2017).

Second and consistent with other forms of research, replicability is a function of (a) the precision with which various steps and decision rules in the process are documented and (b) the transparency with which the rubric is communicated to other researchers. Ideally, researchers should provide datasets (or the code used to harvest and preprocess corpora) and script used to develop and score content. Publication of data and script is commonplace in the NLP literature. Most studies in accounting and finance in contrast do not make data and code available to other researchers to facilitate replication and extension. In addition, a high portion of studies provide little or no information about the specific text preprocessing steps. Examples of good practice include Li (2008), Gruning (2011), Campbell et al. (2013) and Bodnaruk et al. (2014).

---

24 Sometimes authors’ desire for transparency may be overridden by reviewers and editors. For example, the October 2014 version of Lang and Stice-Lawrence (2015) included a detailed discussion of text processing in an appendix, which was subsequently omitted from published version of the paper.
The need for significant researcher judgement combined with opaque reporting practices mean that the proposed benefits of objectivity and replicability may be illusionary for some studies employing computer-based scoring methods. Further, we view this as a potentially more dangerous situation compared with research designs where subjectivity is acknowledged explicitly: in the latter case the warning signs serve to stimulate a degree of healthy scepticism whereas the strong claims of objectivity associated with automated content analysis risk creating a false impression regarding the precision of results and conclusions.

Most automated linguistic analyses refine the raw corpus prior to processing to aid disambiguation and reduce processing time. Common refinements include removing proper nouns and stop words, adjusting hyphenated words, correcting for spelling differences between American English and British English, and converting all words to lower case so that no two same words such as “strategy” and “Strategy” are included in the corpus as different words. The nature of these adjustments is conditional to a large extent on the specific research question and text processing methodology employed.\(^{25}\)

Our review of the literature highlights three concerns relating to preprocessing strategies that compromise objectivity and replicability. First, studies applying broadly similar textual analysis methods such as word counts often apply different preprocessing rules with respect to stemming, stop words, boilerplate text, etc. Inconsistent treatment of the underlying data makes comparisons between studies more difficult than might otherwise be the case. The situation

\(^{25}\) Although stop words dominate frequency counts in many corpora their exclusion is unnecessary in accounting and finance applications that rely on word counts from sentiment categories (LM). Conversely, exclusion of stop words is standard practice in more advanced textual analysis applications such as machine learning in order to reduce dimensionality of the feature space and improve processing efficiency (e.g., Goel et al. 2010). Moreover, stop words such as personal pronouns play a central role in some analyses such as deception detection (Larcker and Zakolyukina 2012, Purda and Skillicorn 2015).
contrasts with studies using accounting and market data where “industry standard” practices on matters such as trimming, scaling and variable definitions have evolved and where extensive sensitivity tests are conducted to assess robustness to alternative approaches.

Second, studies often adopt a particular set of preprocessing rules without explaining the rationale or theory underpinning the choice. In the absence of established norms in the literature, researchers need to justify refinements made to the raw corpus in the context of their research question, the textual analysis methods employed, and the extant textual analysis literature.

Our third observation regarding preprocessing strategies relates to (lack of) transparency. While some studies provide extensive details of the steps involved in cleaning the raw corpus, many offer either scant information on the topic or no material discussion of the issue at all. For example, purging the corpus of proper nouns such as company names, products, countries, individuals, months of the year, etc. is standard practice to reduce the risk of analysing words out of context (LM). While a range of different tools are available to support this task, they can generate different results. Further, since many proper nouns are industry-specific, error levels are likely to depend on sample composition.26 Examples of best practice in this area include Campbell et al. (2013) and Dikolli at al. (2017), where high transparency is defined as provision of sufficient detail to facilitate replication. The lack of precision impedes interpretation and replicability of results (Loughran and McDonald 2016).

5. Assessing the incremental contribution beyond manual analysis

26 For example, words like “crude” and “heavy” describe specific grades of oil in the oil and gas sector; “trust” and “mutual” describe particular organisational types in the financial sector.
This section considers the magnitude of the incremental contribution afforded by automated textual analysis of financial narratives. We evaluate contribution from the perspectives of both the academic research landscape and the professional and regulatory reporting environment. Rather than providing a comprehensive review of extant research studying qualitative financial market information, our goals are (a) to evaluate the research avenues unlocked as a result of applying computational linguistics methods and (b) to compare these contributions against both the insights provided by studies that rely on manual coding methods and the broader financial market debate surrounding the provision and value of qualitative information. We begin by summarizing the current state of play in the accounting and finance literature with respect to the insights provided by automated textual analysis. Next we compare and contrast these insights with those generated by research using manual content analysis methods. Finally, we assess the primary contributions of automated textual analysis research within the context of practical financial market debates (particularly in the financial reporting domain) concerning the provision and role of unstructured data.

5.1 Overview of research using automated textual analysis methods

Tone or sentiment is arguably the most common linguistic property examined using automated methods. In the accounting literature, a significant body of work examines the drivers and consequences of the tone in financial narratives. The main tension in this literature is between the view that narratives are informative for future performance and the view that management use narratives strategically to manipulate users’ perceptions of current and future performance. Consistent with results for other aspects of financial reporting such as discretionary accruals and proforma earnings reporting, findings support both views. On the one hand, Li
(2010) reports that the tone of forward-looking MD&A statements is positively associated with future financial performance, while Feldman et al. (2010) document that MD&A tone changes correlate positively with the filing date market reaction and predict price drift to the subsequent filing date. Similarly, Davies et al. (2012) find that tone predicts for future performance in quarterly earnings press releases.

On the other hand, a number of studies provide evidence consistent with tone management in annual reports, earnings announcements and conference call presentations. Huang et al. (2014) find that abnormal tone in earnings press releases is associated with opportunistic reporting incentives and serves to mislead investors in the short run. Rogers et al. (2011) find that firms with unusually optimistic earnings announcements are more likely to be sued, and that shareholders’ lawsuits specifically target the more optimistic statements. Davis and Tama-Sweet (2012) compare tone in firms’ quarterly press releases with management commentary in the corresponding 10-Q or 10-K and find that the proportion of incremental negative language in the press release is lower for firms that just meet or beat analysts’ earnings forecast. Further and consistent with opportunistic framing, Cho et al. (2010) examine environmental disclosures in section 1 of the 10-K and find that firms with poor environmental performance exhibit more optimism and less certainty. Finally, Allee et al. (2015) measure dispersion of positive and negative tone in the management presentation section of conference calls. Consistent with management seeking to influence investors’ response to company news, they document a tendency for good news to be more dispersed than bad news.

Other research uses tone as a lens to study qualitative information produced by financial intermediaries. A growing body of work examines the informativeness of disclosures provided by analysts. Kothari et al (2009) study the impact of analysts’ reports on firm-level proxies for
information asymmetry and provide inconclusive results. In contrast, Huang et al. (2014) find that the tone of analysts’ reports is incrementally informative for stock price over and above their quantitative outputs, while Caylor et al. (2017) find that report tone has predictive value. Meanwhile, Yukselturk and Tucker (2015) document that report sentiment explains analysts’ quantitative outputs in non-financial crisis periods. Finally, De Franco et al. (2014) examine the tone of the sentences that discuss potential debt-equity conflicts in credit analysts’ reports and conclude that tone is informative for credit spreads and trading volume in the secondary market and offer yields in the primary market.

Work also examines tone in the context of financial media coverage. Kothari et al. (2009) find that the tone of business press disclosures correlate consistently with capital market indicators. Hooghiemstra et al. (2015) examine the effects of media coverage on executive compensation and document a significant positive link between negative CEO press coverage and the level of dissent against executive pay proposals at subsequent AGMs. Findings are consistent either with media coverage causing dissent or with media coverage and dissent being driven by a common unobservable factor.

Readability and understandability are the next most widely studied properties of qualitative disclosures after tone. Loughran and McDonald (2016) provide a detailed review of the literature on readability and how to measure it. Work using a range of readability proxies documents how enhanced disclosure clarity in 10-K filings is associated with more informative reporting and more timely information processing by investors (You and Zhang 2009, Lee 2012). Less sophisticated investors appear to benefit disproportionately from clear and concise reports (Lawrence 2013). Further, Lehavy et al. (2011) find that the value of analysts’ research is inversely related to 10-K readability, and Bonsall and Miller (2017) find that less readable 10-K
narratives are associated with less favourable bond ratings, greater disagreement between rating agencies, and higher credit spreads. Consistent with management actively seeking to strategically obscure information, research documents evidence of lower annual report readability in the presence of poor underlying performance (Li 2008, Lo et al 2017) and excessive CEO remuneration (Hooghiemstra et al. 2017). Regarding the ever-expanding length and potential complexity of 10-K annual reports, Dyer et al (2017) find that increases are largely driven by additional FASB and SEC requirements related to fair values, internal controls and risk factors.

Other work uses a readability lens to study the market impact of conference calls and analyst reports. Consistent with clarity of messaging influencing the informativeness of conference calls, evidence reveals that information asymmetry surrounding the call is positively (negatively) associated with the obfuscation (informativeness) component of the call (Bushee et al. 2018), and that the market response to non-U.S. firms’ conference calls is lower for calls containing greater use of non-plain English and more erroneous expressions (Brochet et al. 2016). In relation to analysts’ reports, De Franco et al. (2015) document a positive link between analyst ability and report readability, while both De Franco et al. (2015) and Hsieh et al. (2016) find the more readable reports are associated with stronger market reactions.

Beyond papers on tone and readability, a large body of research uses textual analysis methods to study the usefulness of qualitative information. A number of studies examine the usefulness of specific aspects financial reporting narratives. Evidence consistent with annual report narratives providing information that is incrementally useful beyond quantitative disclosures includes Hussainey at el. (2003) and Schleicher et al. (2007) in relation to forward-looking management commentary, Merkley (2014) in relation to R&D narratives, Li et al. (2013) in relation to competitive position, and Kravet and Muslu (2013) in relation to risk disclosures.
In addition, Dikolli et al. (2017) demonstrate that language differences between the MD&A and the letter to shareholders provides information about CEO integrity. More generally, Lang and Stice-Lawrence (2015) document a positive association between a proxies for high quality annual reporting and the capital market benefits to transparency. However, Brown and Tucker (2011) predict and find that the usefulness of annual report commentary is restricted to the subset of new content. Finally, a growing body of work also demonstrates that annual report commentary predicts accounting violations and fraud (Goel et al. 2010, Purda and Sillicorn 2015, Hoberg and Lewis 2017) even after controlling for quantitative data, while Larcker and Zakolyukina (2012) document similar results for management commentary in conference calls.

Collective, extant research studying qualitative information using automated processing methods provides new and important insights regarding the causes and capital market consequences of a broad range of unstructured financial market data presented as text. Given the recent explosion in qualitative data (relative to quantitative information), research using computational linguistics methods is helping to move the analysis of narrative information up the research agenda towards parity with the analysis of quantitative data. We view this as a necessary step for an area where text is an enduring and essential feature of the information set.

5.2 Comparison with research using manual approaches to analysing qualitative data

Research examining the properties, determinants and economic consequences of financial narratives has a long history in the accounting literature. Merkl-Davies and Brennan (2007) provide a comprehensive review of research in the area, which at the time of publication relied primarily on manual coding methods applied to samples comprising hundreds rather than thousands of observations. In contrast to the majority of recent work large sample work that
focuses primarily on qualitative data produced by U.S. financial market participants, much of this older body of research is based on corporate disclosures in non-U.S. markets. Accordingly, insights provided by studies using manual coding methods form the primary source of information on qualitative disclosures outside the U.S. despite the recent upsurge in automated text-based research. This is particularly true for work on annual reports where significant differences exist between the standardized 10-K and 10-Q reporting template that U.S. registrants must follow and the less prescriptive reporting practices that apply in many other jurisdictions. Indeed, where innovation in annual reporting practice is concerned, the U.S. lags behind international best practice in areas such as strategy and business models, workforce policies, governance, risk, and emissions. As a consequence, much of what we have learned from large sample disclosure research in recent years does not necessarily supplant extant insights derived from manual analyses. Moreover, many research questions and insights provided by manually-coded studies examining non-U.S. disclosures address issues that cannot be directly studied in the U.S. because comparable disclosures do not currently exist. Examples include non-GAAP disclosures in the annual report (Voulgaris et al. 2014), environmental disclosures (Neua et al. 1998), and the interaction between text, graphs and pictorial content (Davidson 2008, Beattie and Jones 2008).

In addition to work examining aspects of disclosure that have not been studied to date using automated methods, significant work examines readability and obfuscation using manually coded datasets. This work overlaps significantly with work reviewed in section 5.2 using large sample methods. Many of the insights provided by these early manual studies provide insights consistent with those presented as “new” contributions in more recent large sample studies. For example, evidence that annual report readability correlates with poor performance and impairs
usefulness is discussed in Jones and Shoemaker (1994), while Merkl-Davies and Brennan (2007) summarize an extensive body of research providing compelling evidence that management use narrative disclosures to influence readers’ perceptions of financial and non-financial performance. For example, Schleicher and Walker (2010) and Schleicher (2012) score forward-looking content in the outlook sections in a sample of U.K. annual reports and find that firms with impending performance declines are associated with more upward-biased tone, non-specific time horizons, segmental forecasts, and references to aims and objectives unsupported by directional and quantitative forecasts.

Manual coding approaches have also being used to study nuanced aspects of qualitative data generally and managerial opportunism in particular. Examples include Hollander et al. (2010) who find that management often leave questions unanswered during conference calls and that the investors respond negatively to managers failure to respond, and Guillaman-Saorin et al. (2012) who find that Spanish firms use the headlines in their earnings announcements to opportunistically highlight good news and downplay bad news. Work on self-serving attribution bias also provides deeper insights regarding the informative versus strategic reporting roles of corporate narratives. Consistent with the informativeness view, Baginski et al. (2004) find that the decision of management to issue attributions alongside their earnings forecasts does not reflect managerial self-serving opportunism. In contrast, Clatworthy and Jones (2003) report evidence consistent with attribution bias for a sample of U.K. firms’ Chairman’s Statements, while Kimborough and Wang (2014) analyse apparently self-serving attributions U.S. firms’ quarterly earnings announcements and conclude that investors are not fooled by such behaviour.

On balance, we conclude that while large-sample automated methods have superseded manual methods for some research questions relating to narrative disclosures, insights provided
by manual analyses offer significant incremental insights in many other areas. Moreover, some of the new allegedly new contributions based on large-sample automated analysis risk overstating their contribution insofar as they report evidence consistent with conclusions presented in earlier work relying on manual scoring methods.

5.3 Relevance to the professional and regulatory agendas

Placing extant research within the context of professional and regulatory debates concerning the provision and use of unstructured data sheds light on the relevance of extant research. We focus primarily on financial reporting and securities market regulation in this section where automated processing of unstructured qualitative information offers particularly promising opportunities for accounting and financial researchers to contribute to practice.

One of the clearest examples where research is informing practice is in the area of investor protection. There exists compelling evidence that variables derived from qualitative information provide incremental predictive power for fraud and GAAP violations beyond quantitative data (Larcker and Zokolyukina 2012, Purda and Sillcorn 2015, Hoberg and Lewis 2017). Reflecting such insights, the Deputy Director and Chief Economist at the SEC’s Division of Economic and Risk Analysis highlighted in a speech to the American Accounting Association Midwest Regional Meeting the scale of the contributions text analytics is making to its enforcement actions (Bauguess 2016). For example, the SEC has invested heavily in machine learning and NLP technology to measure topic and tonality signals in firms’ filings that can be mapped into established risk factors. Away from fraud detection, the SEC is also using text analytics to measure and assess how emerging growth companies are availing themselves of
JOBS Act provisions. The scalability benefits provided by computational linguistics mean that work in this area is well placed to contribute further to securities market monitoring.

Financial regulators consistently stress the importance of clear, concise, balanced and forward-looking narrative reporting (IASB 2010, Financial Reporting Council 2015). Consistent with this view, insights provided by automated textual analysis research complement and extent prior manual analysis by confirming the incremental importance over the binary disclosure decision of how and where information is communicated. Research studying report length, readability and tone clearly highlights significant statistical and economic associations between these features and capital market consequences reflecting the usefulness of narrative disclosures.

However, while extant results speak to practical financial reporting issues, they risk a relevance problem in the sense that empirical proxies for features such as readability and balance are not necessarily recognizable as such by financial market professionals. For example, regulators are likely to be sceptical about the relevance and reliability of readability metrics such as the fog index given the origins of the measure and naïve way it measures an inherently complex phenomenon via a bag-of-words algorithm. Although the bog index (Bonsall et al. 2017) represents a possible a step forward insofar as it links more directly with use of plain English (SEC 1998), most regulators still struggle to make the connection between this proxy for clarity and their understanding of high quality narrative reporting. By way of illustration, consider the characteristics of ineffective financial statement communication identified in IASB’s Discussion Paper on the principles of disclosure (IASB 2017: 19-20, para 2.4), which include: generic or boilerplate descriptions; use of unclear descriptions and undefined technical jargon; poor organisation of information including failure to provide a contents page or other navigation aids; unclear linkage between related but diffuse pieces of information; unnecessary
duplication of information; disclosing information in a format that is inconsistent with industry practice or changing the way information is disclosed from period to period; using narrative disclosure when a table would be more effective; and omitting material information or including immaterial information that might obscure material content. Similar features are identified by the Financial Reporting Council (FRC) (2015) with respect to annual report disclosures more generally. Precisely how these features map into constructs such as the fog and bog index is unclear, particularly given the absence of manual evaluations concerning the validity of these proxies in a financial reporting context (see section 3.4).

Further, while the majority of automated textual analysis research operates at a relatively aggregate level (e.g., the entire earnings announcement, the entire 10-K, or the complete MD&A section), practitioners tend to be interested in more granular issues such as the format and content of disclosures relating to a particular footnote, the placement of specific disclosures within the overall reporting package (e.g., in the financial statements versus in the narrative section of the annual report versus on the company’s website), and limits on the use of jargon, acronyms and industry-specific terms without clear definitions (FRC 2015). In our view, automated textual analysis research in accounting is currently well short of providing meaningful insights on financial narratives at this micro level; and whether such issues will ever be fully amenable to a computation linguistics treatment remains an open question. Similarly, the IASB’s practice statement on management commentary highlights the following two reporting principles underpinning such disclosures: providing management’s view of the entity’s performance, position and progress; and supplementing and complementing information presented in the financial statements (IASB 2010: 8, para 12). In the IASB’s view, ensuring alignment with these principles requires management commentary to include information that meets the qualitative
characteristics described in the IASB’s conceptual framework (IASB 2010: 8, para 13). While properties such as relevance and timeliness may be amenable to investigation using automated processing methods, other qualitative characteristics such as comparability, understandability and verifiability are likely to prove more challenging to operationalize automatically.

An area of narrative reporting that is attracting growing attention from the practitioners and regulators is integrated reporting and the related issue of business models (International Integrated Reporting Council 2011, European Financial Reporting Advisory Group 2013). Integrated Reporting as a brand aims to bring together material information about an organization’s strategy, governance, performance and prospects in a way that reflects the commercial, social and environmental context within which it operates (International Integrated Reporting Council 2011: 6). More generally, the view that high quality financial reporting should involve a clear narrative that integrates various elements of business performance and practice around issues of sustainability and long-term value creation is a common theme among regulators (IASB 2010, FRC 2015). Automated textual analysis research in accounting is yet to address this issue. While approaches such as measuring the degree of cross-referencing in the text may provide opportunities to contribute to this debate, the potential for developing large sample proxies for the degree of narrative integration and the quality of business model reporting is low in our view. Accordingly, whether computational methods offer the best means for researchers to engage with regulators on such issues remains unclear.

Collectively, the examples discussed above highlight clear opportunities where research using computational linguistics methods to study financial text can contribute significantly to practice. These contributions are most likely to occur where there exists a need to identify specific, clearly defined content or measure relatively high-level textual features (e.g., bias and
volume) in a large corpus of unstructured data. On the other hand, the scope for automated methods to outperform manual inspection when examining more nuanced aspects of narrative reporting quality such as materiality, integration, comparability, understandability and clarity appears limited given that these concepts prove challenging for humans to define and measure consistently. Our primary concern is that the allure of large datasets and the low entry costs associated with constructing measures such as tone and readability could serve to drive a further wedge between research and practice in areas such as financial reporting disclosure.

6. New horizons in textual analysis

This section discusses a series of established methods from the computational linguistics literature that have yet to gain traction in mainstream accounting and finance research but which we believe offer significant potential. For each topic we describe the key method(s), briefly discuss example nonfinancial applications, identify resources to support application, and provide examples of how the method could be usefully applied by accounting and finance researchers.

6.1 Information Extraction

Information extraction (IE) is a subarea of NLP concerned with extracting structured information from unstructured or semi-structured machine readable textual documents (Cowie and Lehnert 1996, Rodrigues and Teixeira 2015). The field enables the use of unstructured, distributed information in a structured way. IE analyses text to help extract information related to events, entities, and relationships to answer questions such as who did what to whom (Marquez et
al. 2008, Badieh and van Keulen 2014). Important subfields where IE and NLP are intensively integrated include Named Entity Recognition and automatic text summarization, which are discussed separately in sections 6.2 and 6.4 below.

IE opens the door for econometric studies based on information contained in user reviews, employee evaluations, surveys, news and other textual sources. This is significant because traditionally economists have been limited to studying only those phenomena where data are available, often involving expensive manual collection or through sheer chance of existing datasets. The ability to work with textual data means that accounting and finance researchers are potentially much less constrained in the topics available for study and also in the empirical constructs used to study traditional topics. Accordingly, IE offers almost unlimited opportunities to accounting and finance research for constructing new datasets and creating new empirical proxies. Examples include: key items of information from the notes to the financial statements (including accounting policies) not reported in standard commercial databases; details of valuation models, cost of capital estimates and growth forecasts from analysts’ reports; aspects of executive pay not captured by standard databases; auditors’ reports; CEO letters to shareholders; deal information from IPO prospectuses and merge announcements; and emissions data from environmental reports.

6.2 Named Entity Recognition

Named Entity Recognition (NER) is an information extraction task associated with isolating and then classifying named entities into predefined categories such as person names,

---

27 Unlike information retrieval which is concerned with identifying relevant documents by the use of a search engine to answer a query, IE finds relevant information by producing structured data ready for post-processing. For example extracting CEO names from annual reports (Manning et al. 2008a).
locations and organizations. The simplest NER method for extracting named entities relies on handcrafted lists such as a gazetteer of all geographical locations, personal names or organization names. For example, Rau (1991) applied a heuristics approach to detecting company names and found that approximately four percent of the words in a one million-word corpus of financial news are constituents of company names. Rau’s (1991) bag-of-words NER method used a set of handcrafted rules to detect whether a certain word refers to a company name. The main problem with this method is that it is hard to provide comprehensive coverage of all possible names.

A more appealing approach is a predictive system that uses statistical inference to detect patterns associated with a particular entity (Stevenson and Gaizauskas, 2000). Accordingly, more recent NER work relies on statistical natural language processing methods used for IE. The main statistical methods used to train an NER system are Hidden Markov Models (HMMs) (Leek, 1997), Conditional Markov Models (CMMs) (Borthwick, 1999) and Conditional Random Fields (CRFs) (Lafferty et al., 2001).

Well-known machine learning NER systems include the Stanford NER (Finkel et al. 2005) and the English NER (Tjong Kim Sang and De Meulder 2003). These methods rely on large volumes of manually annotated training data to achieve high accuracy. For example, the English NER dataset comprises news stories between August 1996 and August 1997 collected from the Reuters Corpus and then manually annotated. Retraining or reannotation is not usually required as these methods learn the information involved in extracting a certain named entity and are therefore able to extract names that do not appear in the original training corpus. Ratinov and Roth (2009) show that the NER systems trained on CONLL 2003 are domain independent.

In a financial context, Chien-Liang et al. (2013) use Stanford NER to extract person names from 10-K filings to aid opinion mining of annual reports. Other example where the
approach could be used productively include: detecting country names in annual reports and earnings announcements to help construct strategic and forward-looking proxies for geographic exposure (as opposed to more restrictive and backward-looking measures based on segmental disclosures); isolating accounting line items and ratios (e.g., EBIT, EBITDA, ROE, EPS) discussed in management commentary, analysts’ reports and the financial media as a means of proxying for critical performance metrics that drive value or may be subject to manipulation; identifying proforma earnings constructs.

6.3 Semantics

The aim here is to make meaning explicit (link back to the motivation for Loughran and McDonald’s work on domain specific lexicons).

<To complete>

6.4 Summarization

The volume of available information, particularly from web-based sources, is increasing rapidly. The demand for systems that automatically summarize documents is therefore also increasing and consequently text summarization has grown rapidly into a major research area. At the conceptual level, text summarization is the process of summarizing a single document or a set of related documents to capture the most important events and present them in the appropriate sequence. Consistent with this view, automatic text summarization is the process of producing a shortened version of a text by the use of computers. The aim is that the summary should convey the key contributions of the text, which means only key sentences should appear in the summary.
Automated text summarization therefore involves identifying key sentences. The process of defining key sentences is highly dependent on the summarization method used.

Two generic approaches to automatic summarization are extractive (Funk et al., 2007; Bawakid, Oussalah, 2008) and abstractive (Ganesan et al., 2010; Genest and Lapalme, 2012). Extractive summarization methods are the dominant automatic text summarization approach. The approach seeks to extract (up to a specified limit) the key sentences or paragraphs from the text and then orders them in a way that yields a coherent summary. Most summarizer systems use sentences rather than larger units such as paragraphs, and the extracted units may differ significantly depending on the summarization algorithm used. Abstractive summarization applies language dependent tools and natural language generation (NLG) technology with the aim of mimicking human summarization methods. Accordingly, these summarizers can include words not present in the original document and tend to be much more challenging to implement.

Both approaches have been used in single- and multi-document text summarization settings. The length of the generated summary can vary and ultimately depends on the purpose of the summarization process. Task descriptions of the summarization tracks organized by the Document Understanding Conference (DUC, http://duc.nist.gov/) and the Text Analysis Conference (TAC, https://tac.nist.gov/) specified the summary length to be within a certain limit of words (e.g., between 240 and 250 words inclusive, white-space-delimited tokens). Summaries over the size limit were truncated and summaries below the size limit were penalized. The compression ratio (i.e., how much shorter the summary is than the original) in other summarizers varies from short summaries (e.g., 100 words) (Angheluta et al. 2004) to very short summaries (e.g., ten words) (Douzidia and Lapalme 2004).
Extraction tends to play an important role in single- and multi-document summarization since the process works on extracting sections of a document or a collection of documents that convey the key contributions of the text.28 Statistical metrics are normally used to assess the importance of extracted units. Early approaches towards text summarization worked on giving each unit a score based on features such as word frequencies (Luhn, 1958), position in the text (Baxendale, 1958), and the presence of key phrases (Edmundson, 1969). The main drawback of these approaches concerns the difficulty of extracting semantically related sentences because the approaches do not capture any linguistic features (i.e. cannot differentiate between nouns and verbs). These limitations motivated the adoption of more sophisticated approaches that are able to consider and extract such features. Machine learning is one approach that is used to identify important features, supported by NLP techniques for identifying key passages and relationships between words and sentences (i.e., semantic similarity) (Kupiec et al. 1995, Jaqua et al. 2004, Leite and Rino 2008). Other approaches include statistical approaches. For example, Fung and Ngai (2006) use Hidden Markov Models to reflect that the probability including a given sentence in summarization extract likely depends on whether the previous sentence has also been included. The advantage of using machine learning and statistical approaches lies in the fact that machine learning approaches can mimic human summarization techniques. Consistent with other supervised machine learning applications, the main drawback of the approach is the cost of creating training data, which usually requires human participants to generate a large number of summaries (McCargar 2004).

28 A key section may take the form of a sentence, paragraph or a number of n-grams.
Clustering algorithms that assign observations into subsets (clusters) have also been used to group relevant sentences together to form a summary (Kruengkrai and Jaruskulchai 2003, Liu and Lindroos 2006). Clustering algorithms are useful when no training data are available but their main drawback is high sensitivity to outliers and noise (Herranz and Martinez 2009).

Cardinaels et al. (2017) is the only paper in accounting and finance of we are aware that uses statistical and heuristic summarizers to generate summaries of financial disclosures. Findings reveal that automatic, algorithm-based summaries of earnings releases are generally less positively biased than management summaries, and that investors who receive an earnings release accompanied by an automatic summary arrive at more conservative (i.e., lower) valuation judgments. Other promising applications of the technology could include: detecting bias in IPO prospectuses, shareholder letters and other important communications with market participants; isolating statistically important sentences within a disclosure and then testing how management order and present that information as a means of either obfuscating or enhancing informativeness; and detecting bias in earnings announcements relative to annual report commentary in the spirit of Davies and Tama-Sweet (2012).

6.5 Text Classification

Text classification, also known as text categorization, is the task of assigning predefined categories to unstructured text. A common application of text classification is sentiment analysis, where the goal is to categorize statements or documents according to their tone. However, text classification represents a much broader concept and has many more applications than merely estimating sentiment. The most common application of the method is in web search engines where the goal is to provide users with relevant information on the search topic of interest. In
such cases, effective disambiguation is critical to the classification task (e.g., differentiating between financial institutions and rivers when a search query includes the word “bank”). Another established application of text classifiers involves classifying academic journals into predefined topic clusters such as computer Science, linguistics, business and management, etc. Meanwhile, the ongoing VarDial workshop series (http://ttg.uni-saarland.de/vardial2017/) focus on automatic classification of languages (e.g., determining features that distinguish English text from Italian) and dialects within a certain language such as distinguishing European from Brazilian Portuguese (Zampieri and Gebre 2012).

Early work on text classification used statistical and Naïve Bayes models with tf–idf measures to rank the importance of words in a document (Kalt and Croft 1996, McCallum et al. 1998). The drawback of this simple approach is that it does not capture the meaning of extracted words and as a result, tf–idf will assign the same weight to “fine” whether it is used in the context of quality or in relation to payments for unlawful act. Recent text classification research has expanded to include more advanced textual features by using NLP and corpus linguistics approaches, including extracting syntactic and semantic features such as POS and semantic tags to assist with disambiguation problems for words that have a different meaning depending on the context (Iyyer et al. 2015, Bloehdorn and Moschitti 2007).

Similar to sentiment analysis, text classification challenges include domain dependencies (He et al., 2011; Kanayama and Nasukawa, 2006) and ambiguities caused by complex languages such as Arabic (Abdulla et al. 2014, Duwairi and Qarqaz 2014). In addition and as text classifiers work with large volumes of textual documents, high dimensionality and sparse data are represent significant challenges (AbuZeina and Al-Anzi 2017).
While text classifiers have been used in the financial literature to measure tone, deception, and financing constrains, many other potentially important applications exist including: text-derived classifications of firms to industrial sectors; peer selection for performance measurement and valuation; identifying CEO traits such as overconfidence; and distinguishing between boilerplate versus informative discussions of features such as risk and governance.

7. Conclusions

We examine the advantages and disadvantages associated with automated analysis of financial market text. We begin by articulating the justifications for not applying computational approaches to study qualitative information in financial markets. This analysis helps pinpoint the nature of the concerns over the value of applying automated textual analysis methods and the reliability of resulting conclusions. Our analysis identifies concern over how the methods are applied and the relevance of the research questions examined as the most credible argument against automated content analysis approaches. We therefore proceed to evaluate extant research in light of this concern using three distinct but complementary lenses.

Our first evaluation lens compares the application of methods in accounting and finance to four core principles that underpin the computational linguistics approach. We conclude that while beacons of best practice exist the accounting and finance literature, many papers fail to follow core principles and as such sceptics are right to question the reliability and economic significance of work in this area. Our second evaluation lens focuses on the proposed advantages of automated textual analysis (over manual coding) identified by Li (2010). These advantages include lowering the costs of scoring textual content, promoting generalizability, minimizing
research subjectivity, improving replicability, and increasing statistical power. Our review of extant research leads us to conclude that these benefits are less pronounced in practice than theory suggest due to a range of implementation issues. First, claims that computational methods economize on the degree of manual coding are only valid for the most basic type of content analysis. Rather, automated methods involve significant researcher subjectivity that is often not discussed transparency, thereby impairing replication. Generalizability is also constrained by a narrow focus on specific types of financial text that are associated with low retrieval costs, and by concerns over the portability of wordlists across corpora and institutional settings. Further, the benefits of enhanced statistical power are also questionable given the degree of measurement error associated with text extraction and classification algorithms. The evidence lends further weight to sceptics’ concerns over the reliance placed on this emerging body of work.

Our third evaluation lens focuses on the magnitude of the incremental contribution offered by automated textual analysis of financial narratives over prior research employing manual content analysis. We document examples where use of computational methods provides significant new insights. However, we also show that conclusions offered by some of the so-called pioneering studies using automated scoring methods risk overstating their contribution insofar as they report evidence consistent with conclusions presented in earlier work relying on manual scoring methods. Further, we show that extant research using manual coding methods offers significant incremental insights that are unlikely to be superseded by studies using automated methods given the complex nature of the research questions and the corresponding demand for deep semantic parsing.

Collectively, our review suggests that while rigorous and transparent application of computational methods provides significant opportunities to further our understanding of
qualitative information in financial markets, extant research risks overstating the magnitude of its incremental contribution relative to manual analysis. A key conclusion emerging from our analysis is that computational methods and high quality manual content analysis represent complementary approaches to studying financial text.
References


view=fit&pagemode=none


66


