INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1386

NATURAL LANGUAGE PROCESSING IN ACCOUNTING, AUDITING AND FINANCE: A SYNTHESIS OF THE LITERATURE WITH A ROADMAP FOR FUTURE RESEARCH

INGRID E. FISHER[a]*, MARGARET R. GARNSEY[b] AND MARK E. HUGHES[a]
[a] *The University at Albany, State University of New York, Albany, NY, USA*
[b] *Siena College, Loudonville, NY, USA*

SUMMARY
Natural language processing (NLP) is a part of the artificial intelligence domain focused on communication between humans and computers. NLP attempts to address the inherent problem that while human communications are often ambiguous and imprecise, computers require unambiguous and precise messages to enable understanding. The accounting, auditing and finance domains frequently put forth textual documents intended to communicate a wide variety of messages, including, but not limited to, corporate financial performance, management's assessment of current and future firm performance, analysts' assessments of firm performance, domain standards and regulations as well as evidence of compliance with relevant standards and regulations. NLP applications have been used to mine these documents to obtain insights, make inferences and to create additional methodologies and artefacts to advance knowledge in accounting, auditing and finance. This paper synthesizes the extant literature in NLP in accounting, auditing and finance to establish the state of current knowledge and to identify paths for future research. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: computational linguistics; natural language processing; text mining

1. INTRODUCTION

1.1. Background and Motivation

Natural language processing (NLP) focuses on the study and facilitation of communication between people and computers. One of the central problems in artificial intelligence (AI) is that of communication. AI is typically defined as intelligence that is demonstrated by software or machinery that imitates the workings of the human mind (Tung, Quek, & Cheng 2004). It is an inherently interdisciplinary field of study that draws from domains as diverse as psychology, computer science, linguistics and neuroscience. NLP, more specifically, encompasses a range of computational techniques for analysing and representing texts at one or more levels of linguistic analysis to enable human-like language processing for a range of particular tasks or applications.

Although NLP has been described in many ways, the working definition underlying this discussion follows the notion that NLP 'is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose

* Correspondence to: Ingrid E. Fisher, The University at Albany, State University of New York, Albany, NY, USA. E-mail: ifisher@albany.edu

of achieving human-like language processing for a range of tasks or applications' (Liddy, 2001). This implies, as demonstrated throughout this paper, that 'there are multiple methods or techniques from which to choose to accomplish a particular type of language analysis' (Liddy, 2001). We reference a variety of analytical methods that incorporate NLP data, at various 'levels of linguistic analysis' (Liddy, 2001). However, since 'NLP is considered a discipline within Artificial Intelligence (AI)' we focus most heavily on the AI-related applications of NLP (Liddy, 2001). Despite the plethora of analytical methods that are associated with and use data produced by NLP, it comprises a unique set of 'computational techniques' that should not be confused or conflated with the many analytical tools referenced in the text that follows.

Fifty years ago Goldberg (1965) stated: 'it is scarcely an exaggeration to say that the problem of communication is the axial problem in accounting'. Accounting practice is replete with written documents intended to communicate such messages as, but not limited to: current accounting standards, past and expected future corporate performance (along multiple dimensions), policies and practices embodied by financial statements and the results of financial statement audits. The growth in digital and social media usage by businesses has further increased the volume of unstructured text documents. Since the dissemination of these documents and messages is now largely automated by and through computers, we believe that NLP research and applications have considerable potential to enhance communication in the areas of accounting, auditing and finance. Our paper is motivated by the desire to determine the state of the extant literature in NLP in accounting, auditing and finance so as to inform and guide future research efforts.

1.2. Objectives

This paper examines the NLP literature in accounting, auditing and finance. Our objectives are: (1) to synthesize the literature and present the lessons to be learned from prior work, (2) to identify unanswered questions that present ripe opportunity for future research and (3) to identify the constraints that are likely to present challenges to future research. We focus our literature review on combinations of NLP with AI. Machine learning (ML) is often considered to be a subfield of AI. However, the large number of studies employing ML, as opposed to other AI methodologies, drove us to acknowledge the ML literature separately, where appropriate.

While we identify the various methods that accounting, audit and finance-related researchers have used in association with NLP, we do not seek to create an artificial classification of NLP-related studies. Rather, we acknowledge the existence of groups of studies that have applied similar NLP analyses, or that sought to accomplish similar tasks by employing NLP, within the fields of accounting, audit and finance. We also note the diverse analytical methods that have been employed in addition to, or in combination with, NLP. Most importantly, we seek to highlight the variety of applications of NLP that have developed and the ever-increasing uses of NLP-generated data.

The remainder of this paper is organized as follows. Section 2 describes the methodology used to identify the relevant literature and provides an overview of the evolution of NLP literature. A synthesis of the NLP literature follows. We discuss NLP in accounting, audit and finance, focusing on two major uses of NLP: classification (Section 3) and prediction (Section 4). Frequently-observed applications and data sources are highlighted, followed by a review of readability studies in accounting, audit and finance. In Section 5 we identify future research opportunities as well as likely obstacles and provide concluding observations in Section 6.

DOI: 10.1002/isaf

2. METHODOLOGY

2.1. Literature Selection

We began our search for relevant literature by scanning the bibliographies of four literature reviews that surveyed NLP developments in the accounting, audit, and finance domains (Fisher *et al.,* 2010; Li, 2010c; Kearney & Liu, 2014; Nassirtoussi *et al.,* 2014). We expanded our exploration by performing keyword searches of databases, including the Association for Computing Machinery (ACM) Digital Library, ProQuest, and the linked databases accessible via EBSCO Host. The primary search terms used included: text analysis, text mining, natural language processing (NLP), ML, artificial neural networks (ANNs), support vector machines (SVMs), expert systems (ESs), artificial intelligence (AI), latent semantic analysis (LSA), content analysis and computerized content analysis. Finally, the bibliographies of retrieved papers were manually scanned, identifying additional sources.

2.2. Literature Assessment

An analysis of the diverse bibliographic sources included in this study demonstrates the ongoing interest in NLP. We assessed a total of 266 monographs, including 192 journal articles, 38 conference papers, 25 working papers, 1 book chapter and 10 doctoral dissertations, as shown in Table I. Since our primary focus is NLP combined with ML and/or AI, we excluded 86 studies that addressed manual text analysis, a precursor to the computer-facilitated NLP often employed today, and 81 involving basic text mining. Excluding the four literature reviews noted above, 95 studies featured NLP combined with and/or augmented by ML and/or AI. The studies featuring co-occurrences of NLP, ML and/or AI come from three primary domains: accounting, audit and finance, represented by 20, 15 and 60 items respectively, as shown in Table II. Since NLP literature is replete with protracted terms, in the interest of brevity we use complete designations, followed by acronyms, only for initial references. Thereafter, acronyms are used. To aid the reader, a glossary of all acronyms and corresponding terms is provided in Appendix A.

2.3. Brief Overview of the Evolution of Natural Language Processing Research

Assessing the literature from a chronological perspective, it is clear that NLP research has evolved, over time, as illustrated in Figure 1. Researchers' original focus was on manual text analysis, a non-computerized approach used to tap the informational value of linguistic patterns in text-based data. The earliest research we found employed manual content analysis in order to assess the readability of corporate reports. It found that the 'general level of reading was difficult, the human interest value dull,

Table I. Literature categorized by bibliographic source

Retrieved literature	Accounting	Audit	Finance	Total
Journal article	72	26	94	192
Conference papers	11	3	24	38
Working papers	2	2	21	25
Book chapters	1	—	—	1
Doctoral dissertations	3	1	6	10
Total articles retrieved	89	32	145	266

DOI: 10.1002/isaf

Table II. Literature categorized by method

Retrieved literature	Accounting	Audit	Finance	Total
Manual text analysis	48	10	28	86
Basic text mining	19	7	55	81
Literature reviews	2		2	4
NLP + ML/AI	20	15	60	95
Total articles retrieved	89	32	145	266

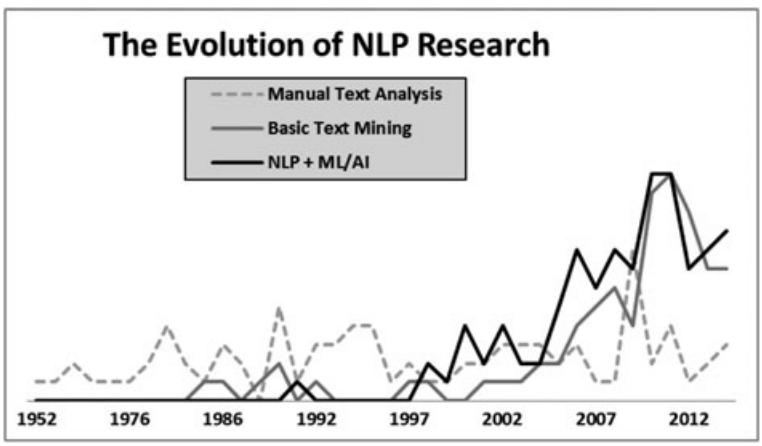

Figure 1. The evolution of NLP research.

and the language was beyond the comprehension of 75% of the U.S. adult population' (Pashalian & Crissy, 1952). In the mid-1980s, studies using basic text mining began to emerge, employing computer programs to extract (i.e. 'mine') relevant textual content. Frazier *et al.*'s (1984) analysis of 74 annual reports featured automated content analysis, using the WORDS text-mining application to identify thematic content associated with management misrepresentation of firm performance. Other than Gangolly *et al.*'s (1991) study, featuring ES, additional studies applying NLP, combined with ML and/or AI, did not emerge until the late 1990s. Thereafter, studies incorporating manual text analysis, text mining and NLP, combined with ML and/or AI, have appeared simultaneously. Figure 1 shows that since 2010 the availability of efficient new text-mining and NLP applications, augmented with ML and/or AI, fuelled an increase in studies using those tools, and a decline in studies involving manual text analysis.

Although this discussion is focused on NLP combined with ML and/or AI, we have identified 86 studies that feature manual text analysis and 81 that employ basic text mining. Since previous reviews summarized these studies, we do not address them here, except where they inform our discussion of NLP. To aid readers who may wish to survey this material, we list the relevant sources in tabular form. Appendix B lists studies featuring manual text analysis. Appendix C lists studies employing basic text mining. Where applicable, we cite literature reviews that have summarized the listed studies.

DOI: 10.1002/isaf

2.4. Analysis of Sample Sizes *N*

After eliminating non-empirical, conceptual manuscripts (e.g. Kamaruddin *et al.,* 2007) and studies where the sample size was unavailable (e.g. Lugmayr & Gossen, 2012), we assessed the comparative sample sizes of the remaining 209 studies. Tracking the largest sample sizes per year, Figure 2 shows the dramatically increased sample sizes associated with the peak period of NLP use (2010–2011).

It seems clear that the trend toward analysing larger data sets has been enabled by the efficiency and enhanced processing capabilities associated with NLP applications combined with ML and/or AI. O'Leary (2013) noted that the age of 'big data', generated by 'the Internet of things' and characterized by oversized data sets, is here, and NLP provides the ability to assess the content and context of text-based data for accounting, audit and finance applications. In the literature we assessed, the seven largest sample sizes were cited in studies between 2010 and 2014. They ranged from approximately 1.8 million to 98.8 million documents. All seven studies employed NLP combined with ML and/or AI, such as SVMs (Bollen *et al.,* 2011a), naive Bayes (NB) (Gilbert & Karahalios, 2010), fuzzy neural networks (FNNs) (Bollen *et al.,* 2011), decision trees (DTs) (Vu, Chang, Ha, & Collier 2012) and ML-based classifiers (Smales, 2014). Four of the seven largest sample sizes (57.14%) were associated with studies employing sentiment analysis (Sinha, 2010; Bollen *et al.,* 2011a; Engelberg *et al.,* 2012; Smales, 2014). All seven studies were finance related.

NLP has enabled researchers to analyse unique datasets associated with digital media, including emerging social media sources, such as Twitter. The analyses of blog content, blogger-created metadata (tags) and other blog-related information sources, such as 'Delicious', have provided rich measures of public sentiment, correlated to predict stock prices, macroeconomic conditions and other phenomena relevant to accounting, audit and finance researchers (O'Leary, 2011). All seven of the studies using large sample sizes, referred to above, were drawn from digital media sources. Bollen *et al.* (2011a, 2011b) and Vu *et al.* (2012) analysed Twitter tweets. Gilbert and Karahalios (2010) assessed 'posts made on the site LiveJournal'. Smales (2014) and Engelberg (2008) used data from the Dow Jones News Service. Sinha (2010) analysed 'firm-specific news items … transmitted to institutional investors over Thomson Reuters' electronic terminals'.

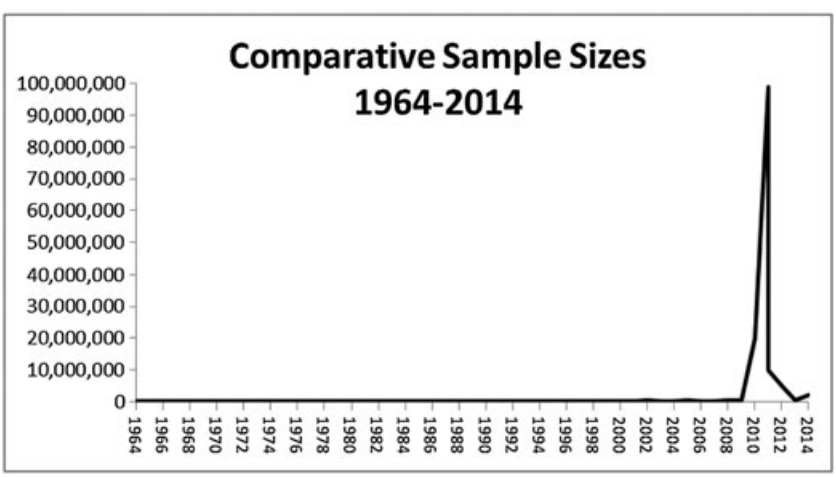

Figure 2. Comparative sample sizes (1964–2014).

DOI: 10.1002/isaf

2.5. Natural Language Processing in Accounting, Audit and Finance

In the accounting, audit and finance literature reviewed, NLP data, analysed using AI algorithms, were typically associated with methodologies such as ESs (Gangolly *et al.,* 1991), intelligent agents (IAs) (Müller *et al.,* 1999), fuzzy neural networks (FNNs) (Bollen *et al.,* 2011a), genetic algorithms (GAs) (Thomas & Sycara, 2002) and self-organizing maps (SOMs) (Magnusson *et al.,* 2005).

NLP combined with ML was represented in algorithmic approaches such as ANNs (Gu *et al.,* 2007), SVMs (Humpherys *et al.,* 2011), DTs (Mahajan *et al.,* 2008), hierarchical agglomerative clustering (HAC) (Garnsey, 2006a; Chakraborty & Vasarhelyi, 2010), keyword in context (KWIC) (Grant & Conlon, 2006); Chinese lexical analysis system (CLAS) (Wang & Guo, 2012), classification and regression trees (CARTs) (Henry, 2006), latent semantic indexing (LSI) (Fisher & Garnsey, 2006), conditional random field (CRF) models (Chen *et al.,* 2011), and various statistical classifiers, such as NB (Li, 2010a), logistic regression (LR) (Henry, 2006), support vector regression (SVR) (Li Q *et al.,* 2014), linguistic inquiry and word count (LIWC) (Larcker & Zakolyukina, 2012), *k*-nearest neighbour (k-NN) (Groth & Muntermann, 2011), character *n*-gram (CNG) (Butler & Keselj, 2009), term frequency-inverse document frequency (TF-IDF) (Peramunetilleke & Wong, 2002), categorical *k*-nearest neighbour (CKNN) (Huang & Li, 2011), the linearized phrase structure (LPS) model (Malo *et al.,* 2014) and kernel methods (Cecchini, 2005). Appendix D lists studies featuring NLP, augmented with AI and/or ML algorithms.

3. CLASSIFICATION

3.1. Knowledge Organization, Categorization and Retrieval

Multiple studies have applied NLP to knowledge organization (KO) and categorization, focusing on financial accounting pronouncements issued by the Financial Accounting Standards Board (FASB) or the Security and Exchange Commission's (SEC's) EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database. Gangolly *et al.* (1991) suggested applying NLP, combined with ES, to the creation of semantic knowledge bases (or trees) for financial accounting standards (in Financial Accounting and Reporting Systems (FARS)). NLP, augmented with LSI, was shown to outperform TF-entropy-based analysis in the classification of accounting concepts from the Generally Accepted Accounting Principles (GAAP) corpus (Garnsey, 2006a), and it was used to track changes in the language of accounting pronouncements (Garnsey, 2006b). Gerdes' EDGAR-Analyzer combined NLP-driven information retrieval with IA, allowing users to retrieve subject-specific corporate data from EDGAR with a 96% reduction in manual processing (Gerde, 2003). Grant and Conlon's EDGAR Extraction System (EES) combined NLP with KWIC and Structured Query Language (SQL) to extract financial statement disclosures from 10-Ks (i.e. annual corporate financial reports) more efficiently and effectively than traditional manual processing (Grant & Conlon, 2006).

3.2. Taxonomy and Thesauri Generation

Müller *et al.* (1999) sought to automate taxonomy creation using an IA (International Business Machine's Intelligent Miner) and a topic categorization tool that employed NLP, augmented with HAC. Gangolly and Wu (2000) and Gangolly and Tam (2000) demonstrated that NLP, using TF-IDF for term weighting, along with principal-components analysis, and an agglomerative nesting algorithm,

DOI: 10.1002/isaf

enabled automatic classification of accounting concepts in the FARS database and pension-related topics in the pension notes of 10-Ks respectively. They suggested that this approach might inform future attempts at 'automatic generation of dictionaries and thesauri' (ibid.). Fisher and Garnsey (2006) used LSI and HAC to isolate 'meaningfully related clusters' from 567 Financial Accounting Standard amendments. These were compared with a 'descriptive taxonomy of change' proposed by Fisher (2004), validating Fisher's taxonomy. Chakraborty and Vasarhelyi (2010) and Chakraborty (2011) processed basic word counts with LSI and HAC to generate pension-related taxonomies from corporate filings (10-Ks), achieving 'an overall success rate of 95% … for the test data set'.

Garnsey and Fisher (2008) built a comparative language model using the Cambridge Statistical Language Modeling (CMU) toolkit, in order to track new terms from the FASB database. They employed a combination of 'object filtering', bigram identification and probability-based frequency analyses available in the toolkit. This approach was also used by Gangolly (2008) and Garnsey (2008, 2009) to identify differences between US GAAP, International Financial Reporting Standards (IFRS), the Accounting Standards Codification (ASC) and 'predecessor pronouncements'. Fisher and Garnsey (2010) generated a 'prototype retrieval thesaurus' by applying NLP analyses to GAAP Section 715: Employee Benefits. Huang and Li (2011) identified 25 types of risk factors from 10-Ks, using a 'text classification algorithm called the multi-label categorical *K*-nearest neighbor (ML-CKNN)', which calculated a 'categorical similarity score for each label', achieving 74.94% accuracy. Boritz *et al.* (2013) applied quadratic discriminant analysis, demonstrating the value of 'advanced searching capabilities, dictionary building tools, and visualization tools, including dendrograms in S+ (from HAC), divisive clustering, and concept maps'. This allowed the team to produce a validated term dictionary, categorizing 387 Sarbanes Oxley Act of 2002 (SOX) Section 404 reports into 14 categories of risk associated with IT weaknesses (Boritz *et al.,* 2013). Malo *et al.* (2014) used an LPS model ('a hybrid of rule-based linguistic models and machine-learning techniques') and a multi-class SVM algorithm to analyse the semantic tone of debt disclosures in 10,000 financial news articles, generating a lexicon of finance phrases and a general model to assess the 'semantic orientations' (positive, negative or neutral sentiment) of finance-related narratives.

3.3. Information Retrieval

Fisher and McEwen (2009) advocated for extensible markup language (XML)-based document type definitions, which would enable more effective automatic NLP-based information retrieval from authoritative accounting literature. Garnsey *et al.* (2009) demonstrated the value of thesauri and search term indices, generated by NLP, augmented with LSI and HAC, for searching the FARS database. Using a tool called Leximancer CAQDAS (computer-aided qualitative data analysis software), Crofts and Bisman (2010) employed the program's concept identification and 'cognitive mapping' functions, examining the 'conflicting and often competing notions' of the use of the term 'accountability' as found in 114 accounting-journal articles. Chakraborty (2011); Chakraborty *et al.,* 2014) used NLP combined with DTs and rule-based algorithms, to process over 2000 journal-article abstracts, yielding 87.27% accurate automatic classification of accounting literature, focusing on the accounting domains, treatments and modes of reasoning addressed in the articles.

3.4. Classifying Financial Statement Content

NLP has been proven useful in classifying financial statement content, with applications in accounting, audit and finance. Back *et al.* (2001) employed SOM to determine whether chief executive officer

DOI: 10.1002/isaf

(CEO) reports were consistent with the audited financials. Preprocessed text was recoded into word, sentence and paragraph vectors, yielding histograms and SOMs, which were used to identify document similarities, addressing the period from 1985 to 1989. Analyses of the text-based data were compared with quantitative analyses of nine financial statement ratios, exposing deviations between CEO reports and audited financials (Back *et al.,* 2001). Kamaruddin *et al.* (2007) proposed employing tagging and syntactic relation analysis, using the link grammar parser, combined with conceptual graphs, to enable auditors to detect deviations and anomalies in financial statements. Assessing CEOs' earnings' reports, Chen, Liu, Chang and Tsai (2011, 2013) applied TD-IDF analysis, enhanced by CRF, to track inconsistencies between financial statement disclosures and actual earnings in 22,780 statements, spanning 1996–2007. Employing the MALLET toolkit, opinion-bearing text from the disclosures was extracted using various CRF models, trained on the multi-perspective question answering corpus, and semantic analysis of this text was subsequently compared with data reported in the financial statements (ibid.). Qiu, Jiang, and Deng (2013) used 'Chinese document modelling' (TF-IDF term weighting), application of the 'Chinese document readability index', SVM binary and multi-class classifiers, and regression analyses to measure the disclosure quality of Chinese financial statements.

4. PREDICTION

4.1. Fraud Prediction and Detection

Multiple researchers have demonstrated approaches wherein NLP was combined with other analysis methods in an effort to predict and detect fraud. NLP data, processed by SVM and kernel method algorithms, outperformed LDA and logit analysis, detecting fraud, bankruptcy and restatements by analysing SEC Accounting and Auditing Enforcement Releases (AAERs) and corporate 10-Ks (Cecchini, 2005; Cecchini *et al.,* 2010). Keila and Skillicorn (2005) used LIWC, followed by singular value decomposition (SVD)-plot analysis, to analyse 289,695 emails, demonstrating that linguistic models detect deceptive emails, typified by a 'reduced frequency of first-person pronouns and exclusive words, and elevated frequency of negative emotion words and action verbs'. Goel (2008) used an SVM classifier algorithm to analyse linguistic features of annual reports, including voice, active versus passive tone, and readability, detecting an association between these features and fraudulent financial statements. By analysing the linguistic characteristics of companies' 10-Ks with the NB-based Rainbow classifier algorithm, Goel *et al.* (2010) identified fraudulent financials with 89.51% accuracy. Humpherys *et al.* (2011) used NLP output, processed by Naïve Bayes and C4.5 classifier algorithms, to classify fraudulent versus non-fraudulent management discussion and analyses (MD&As), attaining 67.3% accuracy. Larcker and Zakolyukina (2012) detected deceptive quarterly-earnings conference calls by employing LIWC to transcripts of conference calls and analysing the results with an SVM-based classification model. Purda and Skillicorn (2014) achieved 89% accuracy at financial-statement fraud detection using NLP to classify 240 Canadian AAERs as truthful or fraudulent. After applying part-of-speech (POS) tagging, word frequency counts and the Q-Tagger tool, 'to identify separate instances of the same word being used differently', the authors applied a Random Forests algorithm to generate a 'rank-ordered list of words' (Purda & Skillicorn, 2014). Thereafter, analyses provided by an SVM algorithm were combined with Dechow *et al.*'s (2011) financial misrepresentation *F*-score to classify the reports as truthful or fraudulent (Purda & Skillicorn, 2014).

DOI: 10.1002/isaf

4.2. Predicting Stock Prices and Market Activity

NLP has been deployed when researchers sought to determine the predictive value of text-based content regarding stock prices, trading volume and exchange rates. We identified 51 studies that focus primarily on financial outcomes, stock valuation and market effects.

NLP has been used to facilitate 'news-based trading', wherein analysts seek to isolate financial news that affects stock prices and/or market activity (Das & Chen, 2007). Wuthrich *et al.* (1998) used NLP data, processed by a *k*-NN learning algorithm to assess real-time web-based financial news, predicting market activity with 43.6% accuracy. Using an NB classifier, Gidofalvi (2001) classified textual data associated with upward-trending and downward-trending stock, establishing significant a predictive relationship ($p = 0.002$) between financial news and stock market activity. Lavrenko *et al.* (2000) applied a Bayesian classifier for the same purpose. Peramunetilleke and Wong (2002) used several term-weighting approaches, including TF-IDF, TF-category discrimination (TF-CDF), Boolean keyword weighting and a rule-based classifier, to establish a correlation between financial headlines and foreign currency exchange rates. Seo *et al.* (2002) used NLP data, processed with various combinations of LSA, HAC, an NB classifier and a weighted-majority voting classifier, to analyse 1239 news articles, with the optimal combination yielding a 79% accurate classification of articles that signalled an increase or decrease in stock prices. Thomas and Sycara (2002) processed text from web-based stock-discussion bulletin boards, analysed the output using the NB-based Rainbow classifier algorithm and multiple runs through a genetic algorithm, and generated significant ($p < 0.0001$) excess returns.

As research involving NLP progressed, additional predictive applications were developed, addressing various financial market phenomena. Mittermayer (2004) established an association between the narrative content of 6602 corporate press releases and stock prices by employing their NewsCATS (News Categorization and Trading System) software, which featured TF-IDF term weighting, enhanced with an unsupervised SVM light classifier. Mittermayer and Knolmayer (2006) followed the same approach, employing (1) various feature sets, such as collection term frequency (CTF), chi-squared (CHI), information gain (IG), and odds ratio (OR), (2) various sized feature sets, (3) various document representations, such as within-document frequency (WDF), inverse document frequency (IDF), Boolean, and WDF × IDF, and (4) various classifiers, such as linear SVM, Rocchio, *k*-NN, nonlinear SVM with a Gauss kernel, nonlinear SVM with a sigmoid kernel, and nonlinear SVM with a polynomial kernel. In developing a proposed automated trading system, Jiang (2005) employed a sentence-disambiguating algorithm, POS tagging and named-entity recognition (NER), enhanced by the RIPPER rule-based classifier algorithm. Jiang (2005) was able to isolate merger and acquisition events in Dow Jones NewsWire articles with 98.3% precision. Boukus and Rosenberg (2006) used NLP (including TF-IDF analysis), combined with LSA (including SVD), to demonstrate that Federal Open Market Committee (FOMC) minutes correlate with 'market expectations of future intermediate- and long-term yields'. Schumaker and Chen (2009) employed the 'Arizona Text Extractor (AzTeK) system's bag of words, noun phrases and NER analyses, combined with 'supervised learning of [an] SVM regression' to classify news articles for use in predicting subsequent stock price changes. Schumaker *et al.* (2012) and Sinha (2010) took a similar approach, inputting sentiment analyses to the classifier. Smales (2014) used Ravenpack's Multi-Classifier for Equities sentiment indicator to confirm the 'significant negative relationship' between 2000–2010 news releases and market volatility, measured by the implied volatility index. Chung (2014) used BizPro software, including NB and LR algorithms, to identify business intelligence factors in news articles that were shown to influence market reactions toward companies.

Researchers have applied NLP approaches to predict stock prices and market activity in various world markets. Cho *et al.* (1998) employed various keyword weighting approaches (e.g. simple

DOI: 10.1002/isaf

weighting, vector weighting with class relevance, vector weighting with class relevance and discrimination, and vector weighting with cluster relevance and discrimination), a clustering algorithm, and a rule-based classifier to establish the predictive value of financial news regarding stock prices on the Hang Seng Index. Fung *et al.* (2002, 2003, 2005) combined feature weighting, an incremental-*k*-mean-based clustering algorithm and an SVM classification algorithm to establish the predictive value of financial news regarding stock prices in Hong Kong.

Three studies used NLP-generated data to identify correlations between financial news content and Indian stock prices. Soni *et al.* (2007) applied concept map visualization and an SVM-based classifier. Mahajan *et al.* (2008) employed a latent Dirichlet allocation (LDA) statistical model for topic extraction, and a trainable stacked classifier, which combined the functions of a DT classifier and a nonlinear SVM sigmoid kernel classifier algorithm. Kumar *et al.* (2012) used TF-IDF term weighting and a trainable SVM-based classifier.

In time, multiple NLP-based approaches were used to explore the predictive value of various international accounting and finance-related text sources. Zhai *et al.* (2007) applied POS tagging and TF-IDF weighting, enhanced by Gaussian-radial-basis-function-kernel and polynomial-kernel supervised SVM classifiers, to confirm correlations between the textual content of financial news articles and stock-price trends in Australia. In Finland, Lugmayr and Gossen (2012) proposed analysing broker newsletters with a German-based sentiment analysis SVM (LUGO Sentiment Indicator) to predict Deutscher Aktienindex German Stock Index (DAX 30) trading activity levels. Groth and Muntermann (2009, 2011) used TF-IDF term weighting, combined with SVM supervised classification algorithms, to identify a significant ($p < 0.01$) correlation between the tone of 423 ad hoc German corporate disclosures and stock-price changes. Hagenau *et al.* (2013) employed bi-normal separation-based feature selection, enhanced by an SVM classifier, to predict stock-price changes signalled by German financial news with 71.8% precision. Argentine and Brazilian currency trends were successfully predicted by Jin *et al.* (2013), using their Forex-Foreteller system, which employed topic clustering, sentiment analysis (based on the Loughran–McDonald and AFINN sentiment-analysis dictionaries) and regression analysis. Li Q *et al.* (2014) predicted Chinese stock prices in numerous industries, with accuracy rates varying between 50 and 65%, by analysing news articles and online discussion-board content with their eMAQT (eMedia-Aware Quantitative Trader). Their eMAQT featured POS tagging, NER, sentiment analysis (based on 'the Chinese Loughran and McDonald's Financial Sentiment Dictionary') and SVR, which applied 'a regression technique to the SVM' output (Li Q *et al.,* 2014). Using Chinese news articles and Hong Kong Exchange market prices, Li X *et al.* (2014) employed TF-IDF term weighting, enhanced with multi-kernel SVM regression, to successfully predict Hong Kong stock prices (mean absolute error ≤ 0.10).

4.3. Firm-Specific Predictions

At the firm level, researchers have used NLP to demonstrate associations between text-based content and various financial phenomena. Employing NLP, combined with mutual information (MI) collocation scores and an SOM visualization model, Magnusson *et al.* (2005) illustrated that changes in the linguistic patterns of Nokia, Ericsson and Motorola's quarterly financial filings (10Qs) predicted changes in their financial status during the subsequent fiscal quarter. Core *et al.* (2008), using a self-constructed sentiment analysis algorithm, showed that the tone of press articles regarding CEO pay is a highly accurate predictor ($p < 0.05$) of firms' CEO compensation and turnover. Hanley and Hoberg (2010) employed NLP (including word content analysis and document similarity analysis), augmented by a vector-based classification model, to conduct sentiment and document similarity analyses on 2043

DOI: 10.1002/isaf

initial-public-offering (IPO) prospectuses. They used ordinary least squares (OLS) regressions to show that the tone and content of IPO prospectuses were predictive of future firm performance (Hanley & Hoberg, 2010).

4.4. Predictive Value of Annual Reports and Disclosures

Corporate annual reports contain abundant textual data ripe for NLP analysis. Manual text analysis established that financial statement disclosures, including MD&As, can be predictive of future firm performance (Singhvi, 1968; Ingram & Frazier, 1980; Kelly-Newton, 1980; Clapham & Schwenk, 1991; Clarkson *et al.,* 1994; Kasznik & Lev, 1995; Smith & Taffler, 2000; Miller, 2002; Shon, 2003; Bhojraj *et al.,* 2004; Kothari *et al.,* 2009; Lawrence, 2011). Text-mining studies also reaffirmed the association between disclosures and future firm performance (Frazier *et al.,* 1984; Hussainey *et al.,* 2003; Beattie *et al.,* 2004; Levine & Smith, 2006; Linsley & Shrives, 2006; Mohan, 2006; Hanley & Hoberg, 2008; Muslu *et al.,* 2010; Durnev & Mangen, 2011; Kravet & Muslu, 2011; Merkley, 2011, 2014; Campbell *et al.,* 2014).

Using NLP, researchers continued to verify the predictive relationship between corporate disclosures and future firm performance. Hoberg and Phillips (2009) employed basic, local and broad cosine text-similarity measures and OLS regression to show that the correlation between '10-K product descriptions', representing firm similarity ('asset complementarity'), and the likelihood of profitable mergers is significant ('>15% of the standard deviation of the dependent variable'). Li (2010a, 2010b) used a 'Naıve Bayesian machine learning [classification] algorithm' to show that changes in the tone of MD&As were 'positively correlated with future [firm] performance' ($p < 0.01$). Balakrishnan *et al.* (2010) used TF-IDF weighting, the FOG index, sentiment and tone analysis, and an SVM classifier algorithm to demonstrate a significant ($p < 0.05$) correlation between corporate disclosures and firm performance. Brown and Tucker (2011), using TF-IDF term weighting, enhanced with a vector space model classifier, identified significant ($p < 0.10$) price reactions to MD&A modifications. Bao and Datta (2014) employed an LDA unsupervised topic model, the variational expectation maximization algorithm and the CKNN classification algorithm with similar results. Qiu *et al.* (2014) used TF-IDF analysis, augmented with a linear SVM multi-class classification algorithm and the LibSVM regression algorithm (SVR) to identify a significant ($p < 0.001$) correlation between mandatory annual report disclosures and future firm performance.

Studies have used NLP to establish the association between corporate disclosures, in the form of earnings press releases (EPRs), and post-release market responses. Henry (2006) employed keyword counts (accomplished with Diction 5.0 software) and thesaurus-based keyword analysis, enhanced with the application of a CART algorithm, to identify text-based variables, improving firm-performance predictions by 5.4%. Engelberg (2008) used typed-dependency parsing (using the Stanford parser), enhanced with tone analysis and regression analysis, to demonstrate that the tone of EPRs is positively correlated ($p < 0.01$) with future firm performance.

4.5. Predictive Value of Web Content

Antweiler and Frank (2004) used document length (word counts) and sentiment analysis (bullish, bearish or neither) to fuel NB (Rainbow) and SVM classification algorithms, showing that the tone of Raging Bull and Yahoo! Finance message-board postings and *Wall Street Journal* articles correlate with stock returns, trading volume and stock volatility (pseudo-$R^2 = 0.203$). Das and Chen (2007) applied sentiment analysis (using General Inquirer) to small-investor message-board postings.

DOI: 10.1002/isaf

Employing a simple-majority voting classifier, the authors used the outputs of naive, vector distance, discriminant, adjective–adverb and Bayesian classifiers to generate a composite sentiment index, classifying the postings as bullish, bearish or neutral. Their analyses revealed a significant correlation ($p < 0.10$) between 'management announcements, press releases, third-party news, and regulatory changes', investor opinions, and stock returns (Das & Chen, 2007). Gu *et al.* (2007) used NLP to analyse over 500,000 internet postings to 'virtual investor communities' (VICs), confirming a significant correlation between the quality of postings and investors' participation in VICs. To determine the quality of postings, NLP output was analysed by six classifier algorithms, including: (1) a lexicon-based classifier, (2) a readability-based classifier, employing a genetic algorithm to classify content based on 'word count, mean word length, and number of unique words', (3) a weighted lexicon classifier, (4) a vector distance classifier, (5) a differential weights lexicon classifier and (6) a simple-majority voting classifier that assessed the other five classifier outputs, identifying postings as signal (i.e. buy, hold or sell), noise or neutral (Gu *et al.,* 2007). Gilbert and Karahalios (2010) analysed 20 million posts from Live Journal, using two sentiment-classifier algorithms to identify 'anxious' posts. 'The most informative 100 word stems' were fed to a 'boosted decision tree' classifier and pre-identified 'anxious' terms, from an earlier Live Journal corpus, and analysed by a 'bagged Complement Naïve Bayes algorithm' (Gilbert & Karahalios, 2010). The results were compiled into an Anxiety Index covering 174 trading days, and then compared with stock prices, using Granger causality analysis (Gilbert & Karahalios, 2010). Yu *et al.* (2013) employed NLP, combined with automated sentiment analysis, based on an NB algorithm, to classify social media postings. They demonstrated that social-media postings were significantly ($p < 0.01$) associated with the equity values of the firms referenced in the postings (Yu *et al.,* 2013).

Researchers have assessed the predictive value of Twitter users' business-related 'tweets'. Bollen *et al.* (2011b) employed NLP, analysing the moods of over 9 million Twitter tweets with Opinion Finder, which measured positive and negative sentence tone, and Google Profile of Mood States, which measured six mood dimensions, including 'Calm, Alert, Sure, Vital, Kind, and Happy'. The resulting data set was subjected to Granger Causality analysis and assessment using a self-organizing fuzzy neural network algorithm, predicting changes in the Dow Jones Industrial Average with 87.6% accuracy (Bollen *et al.,* 2011b). Vu *et al.* (2012) used NLP (including NER, POS tagging, a linear CRF model, and a semantic orientation (SO) algorithm) and ML (the Twitter Sentiment Tool), combined with a DT classifier, to assess the sentiment of over 5 million Twitter tweets. The team demonstrated that the sentiment expressed in such textual data predicted stock-price changes in Apple, Google, Microsoft and Amazon stock, with accuracy ranging from 75% to 82.93% (Vu *et al.,* 2012).

4.6. Natural Language Processing and Readability Studies

In the accounting, audit and finance domains, the readability of financial disclosures, corporate announcements and related text documents has been evaluated in tandem with, and in addition to, methods such as NLP to support a variety of research goals. The 47 readability studies cited in this review illustrate the evolution of NLP-based analyses, but an exhaustive assessment of all readability studies exceeds the scope of this discussion. More comprehensive reviews of readability literature have been authored by Jones and Shoemaker (1994) and Merkl-Davies and Brennan (2007).

Researchers had a plethora of readability measures at their disposal, including 'the Automated Readability Index (ARI), the Flesch Reading Ease formula (FRE), the Linsear-Write readability formula (LWRF), the Gunning–Fog Index (FOG), the Simple Measure of Gobbledygook index (SMOG), the Dale–Chall readability formula (DALE), the Lix readability formula (LIX) and the Rix readability

DOI: 10.1002/isaf

formula (RIX)' (Moffitt & Burns, 2009). The FOG index, the FRE formula, and the plain English rule were used most often to assess narrative text in the accounting, audit and finance literature (Brennan *et al.,* 2009; Loughran & McDonald, 2010).

Accounting and audit literature highlighted significant associations between financial-disclosure readability and current financial performance (Courtis, 1998; Henry, 2008), future financial performance (Butler & Keselj, 2009), earnings quality and persistence (Li, 2008; Lee, 2010) and fraud potential (Moffitt & Burns, 2009; Goel *et al.,* 2010; Humpherys *et al.,* 2011; Othman *et al.,* 2012).

In finance, manual text analysis established that annual report readability is a direct predictor of firm performance, with poor readability associated with poor performance and vice versa (Soper & Dolphin, 1964; Still, 1972; Dolphin & Wagley, 1977; Healy, 1977; Holley & Early, 1980; Lebar, 1982; Courtis, 1986, 1995a, 1995b, 1998; Gibson & Schroeder, 1990; Schroeder & Gibson, 1990; Baker & Kare, 1992; Clatworthy & Jones, 2001, 2003, 2006; Lehavy *et al.,* 2011; Smith & Smith, 1971; Smith & Taffler, 1992a, 1992b). Studies employing basic text mining confirmed this phenomenon (Jones, 1988; Nelson & Pritchard, 2007; Lee, 2010; Loughran and McDonald, 2009, 2010, 2014; Miller, 2010; De Franco *et al.,* 2012; Othman *et al.,* 2012).

NLP has been used to corroborate and enhance the value of readability measures in predicting a firm's financial performance. Butler and Keselj (2009) used NLP, augmented with CNG, and combined with Flesch, Flesch–Kincaid, and Fog index scores to assess readability, enabling the team to forecast upcoming annual stock performance with over 60% accuracy. Finance literature has also identified associations between readability and stock valuations (Kravet & Muslu, 2011), price volatility (Loughran & McDonald, 2009, 2010) and market responses to analysts' reports (De Franco *et al.,* 2012; Twedt & Rees, 2012). Of the 267 monographs assessed for this discussion, 42 included readability analyses. Appendix E identifies these studies, listing the type of NLP associated with basic readability analysis, highlighting, where applicable, any specific 'method used with basic readability analysis'.

Figure 3 illustrates the chronological and methodological distribution of the readability-oriented studies we reviewed. We identified 25 manual text-based studies from 1952 through 2011, wherein readability was assessed. Fifteen studies combining text mining and readability analysis were isolated from 1998 on. We located eight studies, beginning from 2006, wherein readability analysis was combined, and/or co-occurred, with NLP, after which the resulting data were processed with ML/AI algorithms.

The nature and purpose of readability studies changed as NLP-related tools became more sophisticated. Pashalian and Crissy's (1952) study used manual content analysis to emphasize that corporate reports were 'beyond the comprehension of 75% of the U.S. adult population'. Recognizing the limitations of simple readability analysis (Beattie *et al.,* 2004), which is 'only a limited component of the overall picture' (Rutherford, 2005), researchers combined text mining with readability analysis to analyse the implications of fluctuations in the readability of corporate financials and market analysts' reports.

In time, researchers combined readability measures with various NLP approaches, processing the results with ML and/or AI algorithms. Readability analysis (as a proxy for textual complexity) was amalgamated with CART to determine if EPRs were predictive of market reactions (Henry, 2006). Gu *et al.* (2007) analysed the data from readability analysis with DT classifiers to categorize messages posted on virtual-investor sites in order to determine the predictive value of postings to such sites. Processing readability data with an SVM algorithm allowed researchers to achieve 62.81% accuracy in forecasting stock performance, as they sought to achieve automated portfolio optimization (Butler & Keselj, 2009). Goel *et al.* (2010) used readability and other semantic analyses, analysed with SVM and NB algorithms, to differentiate fraudulent from non-fraudulent annual reports and to identify the stages of fraud

DOI: 10.1002/isaf

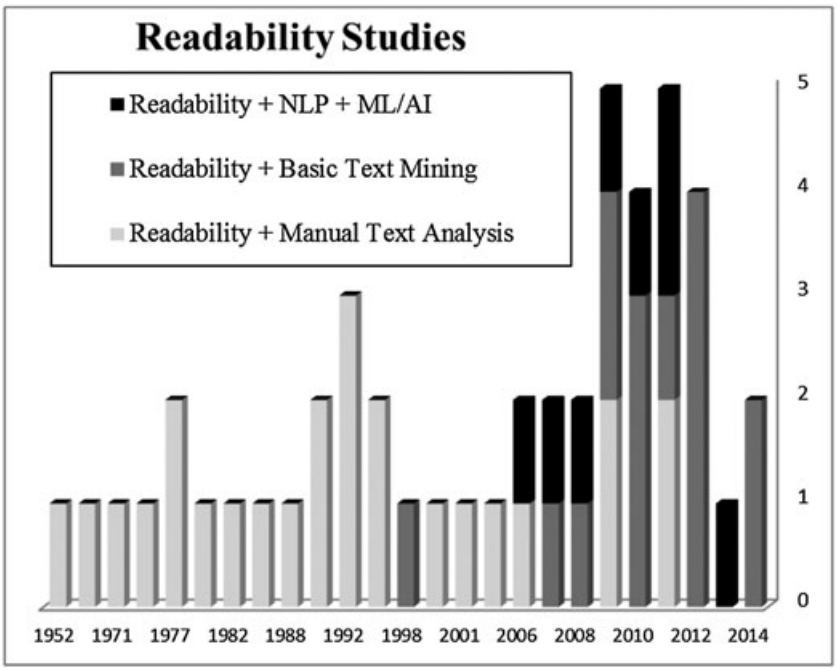

Figure 3. Chronological and methodological distribution of readability studies.

(pre-fraud versus advanced fraud), citing an accuracy rate of 89.51%. Humpherys *et al.* (2011) identified fraudulent financial statements using complexity (i.e. readability) and other linguistic analyses, as calculated by the 'Agent99 Analyzer, a POS tagger, and [a] text analysis tool', followed by analysis of the output with the C4.5 DT, simple NB and SVM algorithms. Qiu *et al.* (2013) combined 'Chinese readability' analysis along with TF-IDF term weighting, processed by SVM algorithms, to evaluate the quality of Chinese annual corporate reports. In a study combining readability and sentiment analyses, Merkley (2014) found 'no evidence that earnings performance related to the readability of R&D disclosures'.

Studies tracked temporal changes in readability, identifying relevant factors that influenced the phenomena. Loughran and McDonald (2010) noted that during the years from 1994 to 2004, 10-K readability declined. Miller (2010) reaffirmed that 'longer and less readable 10-Ks reduce[d] small investor trading volume'. After the SEC launched the 1998 plain English initiative, 'legal jargon and obtuse language' in financial filings decreased, and an improved readability measure, 'Plain English', was used to confirm that 'annual reports with lower earnings' were 'more difficult to read' (Loughran & McDonald, 2010).

Sydserff and Weetman (1999) evaluated 10 UK annual report narratives with 'an alternative to readability formulas', dubbed 'the texture index'. This text-scoring approach measures meaningfulness and comprehensibility of textual content using seven text characteristics, including topicality, intertexuality, conjunction, connectivity, shift in information category, specificity and situationality (Sydserff & Weetman, 1999). Researchers continue to refine readability measures (Loughran & McDonald, 2014). Surprisingly, Loughran and McDonald (2014) demonstrated that '10-K document file size provides a simple readability proxy that outperforms the Fog index', adding that 'an increase in the number of complex words (more than two syllables)' is associated with decreased readability.

DOI: 10.1002/isaf

Will future research provide new readability analyses, perhaps allowing for more nuanced assessments of financial statement usefulness? Will researchers develop new and innovative amalgamations of readability measures, text mining, or NLP, combined with ML and/or AI methods? As Figure 3 demonstrates, the number of readability-based studies peaked between 2009 and 2011, with a marked decline since then. Will the decline persist? Or will the advent of new accounting documents, such as sustainability reports, stimulate additional studies?

5. FUTURE APPLICATIONS OF NATURAL LANGUAGE PROCESSING

Purda and Skillicorn (2014) contend that 'research using textual analysis in accounting and finance is in its infancy'. If this statement is correct, then it behoves accounting, audit and finance researchers to contemplate how to shape future applications of NLP. Although NLP, combined with ML and/or AI, has provided new insights in the accounting, audit and finance domains, additional questions and challenges remain.

5.1. Natural Language Processing in Accounting

In the accounting literature, researchers have suggested various future applications of NLP. Future uses of newly developed parsing tools and text-based classifiers have been proposed, including automated assessments of current journal articles to facilitate taxonomy creation and/or categorization of 'accounting literature by predetermined taxonomy classes' (Chakraborty, 2011). Her research included developing a domain-specific list of keywords, since a formal list was not available. One of the open research questions is to what extent NLP can stand on its own, eliminating the need for manual intervention. In all probability, automated techniques cannot completely replace the role of the human analyst.

Automated exploration of emerging literature to determine 'whether new classes/taxons need to be added' to the existing 'taxons of accounting literature' has been encouraged (Müller *et al.,* 1999; Chakraborty *et al.,* 2014). One significant step that is necessary for this to occur is the comprehensive identification of the relevant documents and the automated collection of those documents.

Automated creation of retrieval thesauri for accounting information has yet to be accomplished (Fisher & Garnsey, 2010). Both thesauri and taxonomies are built from controlled vocabularies. The Master Glossary included with the ASC provides one such vocabulary. This vocabulary is updated periodically (most recently in Accounting Standards Update No. 2014-06) to include new terms, delete terms not used in the ASC and to 'reduce instances of the same term appearing multiple times in the Master Glossary with similar, but not entirely identical, definitions'. Research could explore whether NLP and visualization techniques (such as Leximancer; Crofts & Bisman, 2010) could use that vocabulary and the text in the codification to produce a viable financial accounting thesaurus and/or taxonomy.

5.2. Natural Language Processing in Audit

In the audit literature, investigators have identified several applications of NLP that should be considered in future research. After illustrating the value of NLP-based dendograms and concept maps in analysing the content of auditors' SOX 404 reports, Boritz *et al.* (2013) encouraged future investigators to 'study the efficacy of using automation, rather than manual methods', to code textual evidence and assess the content of MD&As. Cecchini (2005) encouraged future researchers to refine and extend

DOI: 10.1002/isaf

applications of his NLP, SVM and kernel-based model to risk assessments, bankruptcy prediction and fraud detection. Goel *et al.* (2010) envisioned use of NLP, SVM and NB-based classifiers to generate a 'fraud ontology'.

The Public Companies Oversight Board issued Auditing Standard No. 12, Identifying and Assessing Risks of Material Misstatement (PCAOB, 2010). This standard requires auditors of public companies to assess both the external (industry and environmental) risks and internal (company-specific) risks to the possibility of a company materially misstating their financial statements. It would seem that future research could explore the possibility of using NLP to help identify and assess companies' internal risks by examining the characteristics of their reports and disclosures. Given the acknowledgment that many, if not most, companies fail to conduct effective risk assessments combined with the ubiquitous statements regarding the importance of risk assessments (in financial statement audits as well as corporate strategy) (Sklar, 2011), NLP research might be able to make significant contributions to the toolkits of auditors and corporate managers.

An assessment of the readability studies listed in Appendix E reveals that accounting and finance-focused studies exceeded audit-based studies, totalling 16, 20 and 6 respectively. Given the valuable knowledge derived from past accounting and finance-based readability studies, perhaps audit-related readability studies merit further attention.

5.3. Natural Language Processing and Continuous Auditing

NLP, augmented by ML and/or AI, would likely enhance continuous auditing applications. O'Leary (2012) highlighted the benefits of analysing unique information sources, including social media content, such as blogs, wikis and message boards, in order to expand the data assessed in continuous auditing. As noted earlier in this discussion, NLP has proven particularly useful in making such data useful for accounting and audit analyses. Brown *et al.* (2007) advocated for the use of NLP in continuous auditing, highlighting, in particular, the value of combining NLP with neural networks and IA technologies. Neural networks operate well in 'unstructured and data intensive' processing systems, given their ability to cope with 'uncertainty, hidden relationships, and noise', issues commonly faced in continuous auditing systems (Brown *et al.,* 2007). Neural networks are well suited to audit applications, because they 'are particularly effective when a large database of prior examples exists with known inputs and outputs', as is often the case with audit engagements (Brown *et al.,* 2007).

Previous uses of neural networks suggest their potential for continuous auditing. Indeed, neural networks have 'long been used to screen transactions and claims for fraud prevention by banks, credit card companies, and the insurance industry' (Brown *et al.,* 2007).

Additionally, neural networks are able to 'quickly tag unusual transactions for additional screening', which could 'minimize the impact of continuous auditing on the performance of accounting systems' (Brown *et al.,* 2007).

NLP, combined with IAs, would likely be a good fit for continuous audit systems. A unique function associated with IAs is their ability to access 'the internet to seek out information to achieve its specific goal' (Brown *et al.,* 2007). This could be quite useful in a continuous audit system, where an IA could be employed to 'query on-line databases or to query trading partners for information' (Brown *et al.,* 2007).

5.4. Natural Language Processing in Finance

In finance, researchers working on cutting-edge forms of NLP advocated for future studies to build on their accomplishments. Narrative disclosures were identified as untapped repositories of qualitative

DOI: 10.1002/isaf

financial-market data, which could be harnessed to improve investment decisions. Several examples show the untapped potential of textual data, interpreted by means of NLP approaches. After describing their application of NLP, combined with by ANN and SOMs, Back *et al.* (2001) suggested that future research might provide NLP-based tools to empower stakeholders to readily integrate analyses of traditional quantitative data and previously ignored qualitative data. Butler and Keselj (2009) suggested that future studies might use combined readability and CNG analysis to identify 'companies that produce more easily read annual reports, making them … a safer investment'. Das and Chen (2007) posited that future studies could confirm whether their NLP-based classifier algorithm could be used by regulators to detect 'market manipulation' by analysing 'the millions of messages posted to message boards every day'. It remains to be seen what qualitative data sources are yet to be identified and assessed through the application of NLP.

6. CONCLUSIONS

As can be seen from Appendix F, SVMs are by far the most common tools used in NLP-based research. NB is a distant second, followed by hierarchal clustering, statistical methods and TF-IDF weighting. TF-IDF term weighting is also commonly used in the vector space model, which was used in several studies. In information retrieval research, entropy has been found to be superior to other methods, including TF-IDF, in term weighting (Harmon, 1986; Dumais, 1991). It gives higher weights to those terms which are concentrated in a few documents. Research which incorporates term weighting should consider comparing entropy with TF-IDF term weighting to determine if using entropy will significantly impact results.

Loughran and McDonald (2011a) found a majority of terms considered negative in general contexts were not considered negative in accounting/finance contexts. They have developed several domain-specific lists for financial contexts. As mentioned by Kearney and Liu (2014), more research is needed to develop additional authoritative domain-specific lists. This would improve research in the sentiment analysis that was employed by several studies examined.

Natural language remains inherently difficult for computers to understand. Computers require that language be precise, unambiguous and highly structured to effect comprehension. However, natural language is awash with imprecision, ambiguity and relaxed structure. Social context, dialects, humour and slang are among many complicated variables that impact the precision and ambiguity of language. Machine translation stands as one clear example of a specific NLP task that is unlikely to ever be perfected by a computer (Kelly, 2014). Nearly 60 years ago a collaborative effort between Georgetown University and IBM, which produced the 701 Computer (also known as the 'Brain'), promised that machines would be fully translating languages within the next several years (Kelly, 2014). Yet that claim remains unfulfilled. The unique characteristics of natural language continue to present interpretive challenges to automated methods and computers. It is unlikely that these challenges can ever be completely eliminated.

Despite the inherent challenges to NLP, research using NLP has led to certain insights:

- NLP can be used as an effective tool to generate (Chakraborty, 2011) and validate (Fisher & Garnsey, 2006) prototype taxonomies and thesauri (Fisher & Garnsey, 2010).
- NLP can also be used to draw certain inferences:
- Corporate reports were judged as unreadable in 1952 using manual analysis (Pashalian & Crissy, 1952). Today, using NLP, they remain, at least to some degree, difficult to read (Loughran & McDonald, 2010).

DOI: 10.1002/isaf

- The readability of corporate reports and disclosures may be a product of management motivation to obfuscate poor firm performance (Butler & Keselj, 2009) and even fraud (Goel *et al.,* 2010).
- Finance and financial accounting studies have established a consistent relationship between textual characteristics and stock price movement (both upward and downward) (Lavrenko *et al.,* 2000; Gidofalvi, 2001; Seo *et al.,* 2002; Thomas & Sycara, 2002; Schumaker & Chen, 2009; Schumaker *et al.,* 2012).
- NLP holds promise as a tool in fraud detection (Goel *et al.,* 2010).
- Some NLP studies during the peak period of 2010–2012 also examined dramatically larger data sets; however, these studies were all finance-related studies. Given the increased availability and sophistication of digital tools for processing large data sets, it is surprising that it appears that research in accounting and auditing has yet to leverage the power of these tools.

In the areas of taxonomy/thesauri/ontology creation, to date, NLP has only been used to develop prototypes that encompass small subsets of document collections. As but one example, Fisher and Garnsey (2010) present a prototype thesaurus that only addresses a single section of GAAP: Section 715: Employee Benefits. For work in this space to have significant impact, it is necessary for a more comprehensive proof of concept to be completed. The size of the necessary text corpora as well as the distributed location of relevant documents have impeded more comprehensive projects. In addition, the ever-changing nature of the accounting lexicon (Burke, 1997) combined with the inherent difficulties in machine comprehension of natural language raise serious questions as to whether the completion of such a proof of concept is realistic. Possibilities for research combining manual and automated methods are needed. For example, the XBRL taxonomy provides structure that could possibly be used as a starting point for constructing a financial accounting ontology. Wächter *et al.* (2011) have experimented using DOG4DAG to semi-automatically extend an ontology using PubMed extracts.

Despite the fact that some human involvement may be necessary for effective NLP applications, this should not deter future researchers from pursuing advancements in NLP applications in accounting, auditing and finance. A proliferation of accounting documents are available electronically, including accounting pronouncements, SEC regulations, tax laws, tax court cases, comment letters to the FASB, accounting research articles and working papers, accounting pedagogical material (syllabi, case studies, program descriptions) and documents published on corporate websites. The documentation available for analysis is only likely to increase in a business environment characterized by public scrutiny, increased regulation and a focus on compliance. The increased availability of digital documents in accounting, auditing and finance combined with the increased availability and sophistication of digital tools for examining these documents presents an opportunity for fruitful future research, as illustrated in Appendix G, which summarizes all 262 studies (86 manual, 81 basic text-mining, and 95 NLP studies augmented with ML and/or AI) by data source and analysis method. NLP applications offer the means through which to obtain additional insights and create new methods for mining the many textual communications in accounting, auditing and finance.

DOI: 10.1002/isaf

APPENDIX A. GLOSSARY OF NLP-RELATED TERMS

Glossary

Acronym	Complete Designation
10Q	Quarterly Financial Filings (SEC)
AAER	Accounting and Auditing Enforcement Releases (SEC)
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ARI	Automated Readability Index
ASC	Accounting Standards Codification
ATRANS	Automatic Funds Transfer Telex Reader
AzTeK	Arizona Text Extractor
BNS	Bi-normal Separation
BPN	Backpropagation
CAQDAS	Computer-Aided Qualitative Data Analysis Software
CART	Classification and Regression Trees
CKIP	Chinese Knowledge and Information Processing
CKNN	Categorical K-Nearest Neighbor
CHI	Chi-Squared
CLAS	Chinese Lexical Analysis System
CLN	Categorical Learning Network
CMU	Cambridge Statistical Language Modeling Toolkit
CNG	Character N-Gram
CTF	Collection Term Frequency
CRF	Conditional Random Field
DALE	Dale-Chall Readability (formula)
DAX	Deutscher Aktienindex (German) Stock Index
DSA	Document Structure Analysis
DTD	Document Type Definitions
DT	Decision Tree
DWLC	Differential-Weights Lexicon Classifier
ECRS	Executive Compensation Retrieval System
EDGAR	Electronic Data Gathering, Analysis and Retrieval database (SEC)
EES	EDGAR Extraction System
EPR	Earnings Press Releases
ES	Expert Systems
FARS	Financial Accounting and Reporting System
FAS	Financial Accounting Standard
FASB	Financial Accounting Standards Board
FNN	Fuzzy Neural Network
FOG	Gunning Fog Readability Index
FOMC	Federal Open Market Committee
FRAANK	Financial Reporting and Auditing Agent with Net Knowledge

(*Continues*)

DOI: 10.1002/isaf

Appendix (Continued)

Glossary

Acronym	Complete Designation
FRE	Flesch Reading Ease (formula)
GA	Genetic Algorithm
GAAP	Generally Accepted Accounting Principles
GI	General Inquirer
GPOMS	Google Profile of Mood States
HAC	Hierarchical Agglomerative Clustering
IG	Information Gain
IA	Intelligent Agent
IFRS	International Financial Reporting Standards
IPO	Initial Public Offering
KEA	Knowledge Engineering Approach
k-NN	K-Nearest Neighbor
KO	Knowledge Organization
KWIC	Keyword in Context
LBC	Lexicon-Based Classifier
LDA	Latent Dirichlet Allocation/Linear Discriminant Analysis
LIWC	Linguistic Inquiry & Word Count
LIX	Lix (readability formula)
LMWL	Loughran and McDonald Word List
LPS	Linearized Phrase Structure
LR	Logistic Regression
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
LWRF	Linsear-Write Readability Formula
MCQ	Multi-Classifier for Equities
MDA	Multiple-Discriminant Analysis
MD&A	Management Discussion and Analysis
MKSVR	Multi-Kernel Support Vector Regression
ML	Machine Learning
MLR	Multinomial Logistic Regression
MPQA	Multi-Perspective Question Answering
NB	Naïve Bayes
NER	Named-Entity Recognition
NewsCATS	News Categorization and Trading System
NLP	Natural Language Processing
NLTK	Natural Language Processing Toolkit
NUDIST	QDA Software
OR	Odd's Ratio
OLS	Ordinary Least Squares
POS	Part of Speech Tagging

(*Continues*)

DOI: 10.1002/isaf

Appendix (Continued)

Glossary

Acronym	Complete Designation
QDA	Quadratic Discriminant Analysis
RBC	Readability-Based Classifier
RIX	Rix (readability formula)
SCISOR	System for Conceptual Information Summarization, Organization, and Retrieval
SEC	Security and Exchange Commission
SMOG	Simple Measure of Gobbledygook (Index)
SO	Semantic Orientation
SOFNN	Self-Organizing Fuzzy Neural Network
SOM	Self-Organizing Maps
SOX	Sarbanes Oxley Act of 2002
SQL	Structured Query Language
SSE	Shanghai Stock Exchange
SVD	Singular Value Decomposition Vector
SVM	Support Vector Machines
SVR	Support Vector Regression
TDP	Typed Dependency Parsing
TF-CDF	Term Frequency - Category Discrimination
TF-IDF	Term Frequency - Inverse Document Frequency
TST	Twitter Sentiment Tool
VDC	Vector Distance Classifier
VEM	Variational Expectation Maximization
VIC	Virtual Investor Community
VPRS	Variable-Precision Rough Sets
VSM	Vector Space Model
WLC	Weighted-Lexicon Classifier
WDF	Within-Document Frequency
WORDS	Text-mining Software
XML	Extensible Markup Language

DOI: 10.1002/isaf

APPENDIX B: STUDIES EMPLOYING MANUAL TEXT ANALYSIS

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
1	Abrahamson & Amir, 1996	1	JA	Sentiment Analysis	Finance	1,355	US
2	Abrahamson & Park, 1994	1	JA	Sentiment Analysis	Finance	1,000	US
3	Aerts, 1994	1	JA	Content analysis	Acctg	NA	US
4	Badua, 2005	–	DD	Content analysis	Acctg	NA	US
5	Baginski *et al.,* 2004	1	JA	Content analysis	Acctg	951	US
6	Baker & Kare, 1992	1	JA	Readability analysis	Acctg	NA	US
7	Bhojraj *et al.,* 2004	1	JA	Content analysis	Acctg	NA	US
8	Boo & Simnett, 2002	1	JA	Content analysis	Acctg	140	Singapore
9	Bowman, 1984	1	JA	Content analysis	Finance	26	US
10	Brennan *et al.,* 2009	1	JA	Readability analysis	Acctg	NA	US
11	Bricker *et al.,* 1995	1	JA	Content analysis	Acctg	479	US
12	Bryan, 1997	1,3	JA	Content analysis	Acctg	250	US
13	Callahan & Smith, 2004	3	CP	Content analysis	Finance	70	US
14	Churyk *et al.,* 2009	–	JA	Content analysis	Acctg	NA	US
15	Clapham & Schwenk, 1991	1	JA	Content analysis	Acctg	100	US
16	Clarkson *et al.,* 1999	1	JA	Content analysis	Acctg	55	Canada
17	Clarkson *et al.,* 1994	1	JA	Content analysis	Acctg	905	Canada
18	Clatworthy & Jones, 2001	1	JA	Readability analysis	Acctg	60	UK
19	Clatworthy & Jones, 2003	1	JA	Readability analysis	Acctg	100	UK
20	Clatworthy & Jones, 2006	1	JA	Readability analysis	Acctg	100	UK
21	Cohen *et al.,* 2010	1	JA	Content analysis	Audit	39	US
22	Cole, 1990	1	JA	Content analysis	Acctg	100	US
23	Courtis, 1986	1	JA	Readability analysis	Acctg	142	Hong Kong

(*Continues*)

DOI: 10.1002/isaf

Appendix B: (Continued)

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
24	Courtis, 1995a	1	JA	Readability analysis	Acctg	32	Hong Kong
25	Courtis, 1995b	1	JA	Readability analysis	Acctg	NA	Hong Kong
26	D'Aveni & MacMillan, 1990	1	JA	Content analysis	Acctg	114	US
27	Dechow *et al.,* 2009	–	JA	Content analysis	Audit	2,190	US
28	Deumes, 2008	1	JA	Content analysis	Finance	90	Netherlands
29	Dolphin & Wagley, 1977	1	JA	Readability analysis	Finance	NA	NA
30	Donelson *et al.,* 2012	–	JA	Content analysis	Acctg	189	US
31	Farrell & Cobbin, 2000	1	JA	Content analysis	Acctg	57	Australia
32	Folsom *et al.,* 2013	–	WP	Content analysis	Acctg	35,214	US
33	Francis *et al.,* 2002	1	JA	Content analysis	Finance	2,190	US
34	Garcia, 2014	2	JA	Sentiment Analysis	Finance	55,307	US
35	Gibson & Schroeder, 1990	1	JA	Readability analysis	Acctg	40	US
36	Govindarajan, 1980	1	JA	Content analysis	Finance	NA	US
37	Healy, 1977	1	JA	Readability analysis	Acctg	NA	US
38	Hofstedt, 1976	–	JA	Content analysis	Acctg	NA	US
39	Holley & Early, 1980	1	JA	Readability analysis	Acctg	NA	US
40	Hooghiemstra, 2010	–	JA	Sentiment Analysis	Acctg	NA	Netherlands
41	Hooks & Moon, 1993	1	JA	Sentiment Analysis	Acctg	90	US
42	Hoskin *et al.,* 1986	1	JA	Sentiment Analysis	Acctg	676	US
43	Huang, 2011	–	DD	Sentiment Analysis	Finance	12,562	US
44	Huang *et al.,* 2014	2	JA	Sentiment Analysis	Finance	14,475	US
45	Ijiri *et al.,* 1968	–	JA	Sentiment Analysis	Acctg	NA	US
46	Ingram & Frazier, 1980	1	JA	Content analysis	Acctg	40	US

(*Continues*)

DOI: 10.1002/isaf

Appendix B: (Continued)

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
47	Jarvenpaa & Ives, 1990	1	JA	Content analysis	Acctg	649	US
48	Jones, 1988	1	JA	Readability analysis	Acctg	32	UK
49	Kasznik & Lev, 1995	3	JA	Content analysis	Finance	565	US
50	Kelly-Newton, 1980	1	JA	Content analysis	Acctg	NA	US
51	Kimbrough & Wang, 2014	–	JA	Sentiment Analysis	Finance	94	US
52	Klamm & Watson, 2009	–	JA	Sentiment Analysis	Audit	9,272	US
53	Kothari *et al.,* 2009	2,3	JA	Sentiment Analysis	Finance	326,357	US
54	Lawrence, 2011	3	JA	Readability analysis	Finance	91,228	US
55	Lebar, 1982	–	JA	Readability analysis	Acctg	10	US
56	Lehavy *et al.,* 2011	2,3	JA	Readability analysis	Finance	33,704	US
57	Li, 2006	2	DD	Sentiment Analysis	Finance	34,180	US
58	Linsley & Shrives, 2006	–	JA	Content analysis	Audit	79	UK
59	Loughran & McDonald, 2009	–	JA	Readability analysis	Audit	67,044	US
60	Luo *et al.,* 2013	–	JA	Sentiment Analysis	Finance	4,518	US
61	Masli *et al.,* 2009	–	CP	Content analysis	Audit	205	US
62	McConnell *et al.,* 1986	1	JA	Content analysis	Acctg	40	US
63	Meyer & Rigsby, 2001	–	JA	Content analysis	Acctg	NA	US
64	Miller, 2002	1	JA	Content analysis	Acctg	80	US
65	Pashalian & Crissy, 1952	1	JA	Readability analysis	Acctg	NA	NA
66	Pava & Epstein, 1993	1	JA	Content analysis	Finance	25	US
67	Previts & Brown, 1993	1	JA	Content analysis	Acctg	NA	US
68	Previts *et al.,* 1994	1	JA	Content analysis	Finance	479	US

(*Continues*)

DOI: 10.1002/isaf

Appendix B: (Continued)

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
69	Rezaee *et al.,* 2003	1	JA	Content analysis	Audit	94	US
70	Rogers & Grant, 1997	–	JA	Content analysis	Finance	187	US
71	Rutherford, 2005	–	JA	Content analysis	Acctg	44	UK
72	Schroeder & Gibson, 1990	1	JA	Readability analysis	Acctg	40	US
73	Shon, 2003	1	WP	Content analysis	Finance	100	US
74	Singhvi, 1968	1	JA	Sentiment analysis	Finance	200	US
75	Smith & Smith, 1971	1	JA	Readability analysis	Acctg	50	US
76	Smith & Taffler, 1992a	1	JA	Readability analysis	Acctg	66	Australia
77	Smith & Taffler, 1992b	1	JA	Content analysis	Acctg	NA	Australia
78	Smith & Taffler, 2000	1	JA	Content analysis	Acctg	33	Australia
79	Soper & Dolphin, 1964	1	JA	Readability analysis	Acctg	25	US
80	Steele, 1982	1	JA	Content analysis	Finance	100	UK
81	Still, 1972	1	JA	Readability analysis	Finance	NA	US
82	Summers & Sweeney, 1998	–	JA	Content analysis	Audit	51	US
83	Swales, 1988	1	JA	Sentiment analysis	Finance	98	US
84	Sydserff & Weetman, 1999	–	JA	Content analysis	Audit	10	US
85	Yen, Hurst, & Hopkins, 2007	1	JA	Content analysis	Acctg	278	US
86	You & Zhang, 2009	1	JA	Sentiment analysis	Finance	24,259	US

N: sample size; NA: not available or not applicable.

[a]Key to summary references: (1) Fisher *et al.* (2010); (2) Kearney and Liu (2014); (3) Li (2010c); (4) Nassirtoussi *et al.* (2014).

[b]Literature types: journal article (JA), conference paper (CP), working paper (WP), book chapter (BC) and doctoral dissertation (DD).

DOI: 10.1002/isaf

APPENDIX C: STUDIES EMPLOYING BASIC TEXT-MINING

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
1	Andersen *et al.,* 1992	-	CP	JASPER; Rule-Based Pattern Matching	Finance	1,047	US
2	Antweiler & Frank, 2006	1	JA	Naïve Bayes classifier (RAINBOW)	Finance	250,000	US
3	Armesto *et al.,* 2009	-	JA	Diction 5.0 Sentiment analysis	Finance	NA	US
4	Beattie *et al.,* 2004	-	JA	NUDIST Content analysis	Audit	NA	Scotland
5	Beynon *et al.,* 2005	1	JA	VPRS	Acctg	NA	US
6	Bligh & Hess, 2007	-	JA	Diction 5.0 Sentiment analysis	Finance	194	US
7	Bollen, Mao & Pepe 2011	-	CP	Sentiment analysis	Finance	98,853,498	US
8	Born *et al.,* 2013	-	JA	Diction 5.0 Sentiment analysis	Finance	768	Germany
9	Bovee *et al.,* 2005	-	JA	FRAANK	Acctg	50	US
10	Burns *et al.,* 2010	-	WP	Lexical bundles	Audit	202	US
11	Campbell *et al.,* 2014	-	JA	Sentiment analysis	Acctg	9,272	US
12	Carretta *et al.,* 2011	2	WP	General Inquirer (GI)	Finance	16	Italy
13	Chen *et al.,* 2014	2	JA	Sentiment Analysis LMWL	Finance	459,679	Hong Kong
14	Chiu, 2013	-	DD	Content analysis	Acctg	519	US
15	Cong *et al.,* 2007	1	JA	DSA; KEA	Acctg	74,132	US
16	Courtis, 1998	1	JA	Readability Analysis	Audit	120	Hong Kong
117	Craig *et al.,* 2013	-	JA	Diction 5.0 Sentiment analysis	Audit	5	Australia
18	Davis & Tama-Sweet, 2012	2, 3	JA	Diction 5.0 Sentiment analysis	Finance	23,017	US
19	Davis *et al.,* 2014	2	CP	Diction 5.0 Sentiment analysis	Finance	243	US
20	Davis *et al.,* 2006	-	WP	Diction 5.0 Sentiment analysis	Finance	24,000	US
21	Davis *et al.,* 2012	1, 2, 3	JA	Diction 5.0 Sentiment analysis	Finance	24,000	US
22	De Franco *et al.,* 2012	2	JA	Readability Analysis FOG and Flesch	Finance	356,463	Canada

(*Continues*)

DOI: 10.1002/isaf

'Appendix C: (Continued)

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
23	Demers & Vega, 2010	1, 2, 3	WP	GI; Diction 6.0	Finance	20,000	US
24	Ding & Chen, 2006	1	CP	ECRS	Acctg	323	US
25	Doran *et al.,* 2010	2	JA	GI	Finance	233	US
26	Durnev & Mangen, 2011	2	WP	GI; Diction 6.0	Finance	1,905	US
27	Engelberg *et al.,* 2012	-	JA	GI; Sentiment Analysis	Finance	1,888,868	US
28	Feldman *et al.,* 2008	1, 2, 3	WP	GI; Sentiment Analysis	Finance	170,056	US
29	Feldman *et al.,* 2010	-	JA	Sentiment Analysis LMWL	Finance	153,988	US
30	Ferguson, 1997	-	CP	Prolog; Parsing	Acctg	NA	US
31	Ferguson *et al.,* 2014	2	CP	Sentiment analysis	Finance	NA	US
32	Ferris *et al.,* 2013	2	JA	Diction 5.0	Finance	1,175	US
33	Fisher, 2004	-	JA	XML DTDs	Acctg	NA	US
34	Fisher, 2007	-	JA	XML DTDs	Acctg	500	US
35	Frazier *et al.,* 1984	1	JA	WORDS	Acctg	74	US
36	Gangolly, 1989	-	BC	Prolog	Acctg	NA	US
37	Garnsey & Hotaling, 2011	-	JA	NA	Acctg	NA	US
38	Hanley & Hoberg, 2008	-	WP	GI: Content analysis Document Similarity	Finance	2,043	US
39	Henry, 2008	-	JA	TF-IDF	Finance	1,366	US
40	Henry & Leone, 2010	-	WP	TF-IDF	Finance	29,712	US
41	Huang *et al.* 2010	-	JA	CKIP	Finance	12,380	Taiwan
42	Hussainey *et al.,* 2003	1	JA	NUDIST	Finance	60	UK
43	Jacobs & Rau, 1990	-	JA	SCISOR	Finance	729	US
44	Jegadeesh & Wu, 2013	2	JA	Sentiment Analysis LMWL	Finance	45,860	US
45	Kravet & Muslu, 2011	3	JA	Parsing; Readability analysis	Finance	4,315	US
46	Lee, 2010	3	JA	Readability analysis; FOG	Finance	60,161	Taiwan
47	Leinemann *et al.,* 2001	1	JA	Edgar2xml	Acctg	NA	South Africa
48	Levine & Smith, 2006	3	JA	Keyword Search	Acctg	5,983	US
49	Li, 2008	1, 3	JA	Readability analysis; FOG	Acctg	55,719	US

(*Continues*)

DOI: 10.1002/isaf

Appendix C: (Continued)

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
50	Loughran & McDonald, 2009	1	WP	Readability Analysis FOG and Flesch	Finance	42,357	US
51	Loughran & McDonald, 2010	3	WP	Readability Analysis FOG, Flesch, and Plain English	Finance	42,357	US
52	Loughran & McDonald, 2011a	3	JA	Sentiment analysis TF-IDF	Acctg	8,341	US
53	Loughran & McDonald, 2011b	2	JA	GI; Zipf's Law	Finance	50,115	US
54	Loughran & McDonald 2013b	2	JA	Sentiment analysis	Finance	1,887	US
55	Loughran & McDonald, 2014	2	JA	Readability Analysis FOG	Finance	66,707	US
56	Lytinen & Gershman, 1986	-	CP	ATRANS	Finance	NA	US
57	Mendonca, 2009	-	DD	Diction 5.0 Sentiment analysis	Finance	NA	US
58	Merkley, 2011	3	DD	Content analysis	Finance	22,445	US
59	Merkley, 2014	-	JA	Sentiment analysis; Readability Analysis; FOG	Finance	22,482	US
60	Miller, 2010	3	JA	Readability analysis; FOG	Finance	12,771	US
61	Mirakur, 2011	-	JA	Content analysis	Finance	122	US
62	Moffitt & Burns, 2009	-	CP	Readability Analysis; Content analysis; NLTK	Audit	720	US
63	Moffitt & Burns, 2011	-	WP	Python-enabled NLP	Audit	202	US
64	Muslu *et al.,* 2010	3	WP	Keyword Search	Acctg	44,708	US
65	Nelson & Pritchard, 2007	1, 3	CP	Readability analysis; FOG	Acctg	53,315	US
66	Othman *et al.,* 2012	-	CP	Readability analysis Flesch	Finance	16	Malaysia
67	Ozik & Sadka 2012	2	WP	Sentiment analysis; GI	Finance	3,600	US
68	Price *et al.,* 2012	2	JA	Sentiment analysis; GI	Finance	2,800	US
69	Rees & Twedt, 2012	2	JA	Readability Analysis; Sentiment analysis; GI	Finance	2,057	US
70	Rogers *et al.,* 2011	2, 3	JA	Diction 5.0 Sentiment analysis	Finance	165	US

(*Continues*)

DOI: 10.1002/isaf

Appendix C: (Continued)

No.	Source	Summary	Lit Type	Method	Domain	N	Origin
71	Sadique *et al.,* 2013	-	JA	Sentiment analysis; GI	Finance	NA	Malaysia
72	Sadique *et al.,* 2008	1	WP	Sentiment analysis; GI	Finance	1,231	Australia
73	Solomon *et al.,* 2014	-	JA	Sentiment analysis	Finance	1,731	US
74	Sydserff & Weetman, 2002	-	JA	Diction 5.0 Transitivity Index	Audit	26	UK
75	Tennyson *et al.,* 1990	1	JA	WORDS	Finance	46	US
76	Tetlock, 2007	1, 2	JA	GI	GI	NA	US
77	Tetlock, 2011	-	JA	Text similarity	Finance	10,187	US
78	Tetlock *et al.,* 2008	1	JA	GI	Finance	350,000	US
79	Twedt & Rees, 2012	-	JA	Sentiment analysis; GI Readability analysis; FOG	Finance	2,057	US
80	Wang & Guo, 2012	-	JA	CLAS	Finance	50	China
81	Wang, Li, & Cao 2012	-	JA	Wordsmith Tools 4.0	Acctg	120	China

N: sample size; NA: not available or not applicable.

[a]Key to summary references: (1) Fisher *et al.* (2010); (2) Kearney and Liu (2014); (3) Li (2010c); (4) Nassirtoussi *et al.* (2014).

[b]Literature types: journal article (JA), conference paper (CP), working paper (WP), book chapter (BC) and doctoral dissertation (DD).

DOI: 10.1002/isaf

APPENDIX D: STUDIES EMPLOYING NLP COMBINED WITH ML/AI

No.	Source	Lit Type	AI or ML	Method	Domain	N	Data Source	Origin
1	Antweiler & Frank, 2004	JA	ML-SVM	SVM Naïve Bayes	Finance	250,000	Stock Message Boards	Canada
2	Back *et al.,* 2001	JA	AI-SOM	Neural Networks SOMs	Audit	234	Green Gold Financial Reports' DB	Finland
3	Balakrishnan *et al.,* 2010	JA	ML-SVM	SVM	Finance	4,280	1997-2002 Annual Reports	US
4	Bao & Datta, 2014	JA	ML-LDA	LDA	Finance	7,679	2006-2010 Annual Reports	Singapore
5	Bollen, Mao, and Zeng 2011	CP	ML-ANN	Fuzzy Neural Network	Finance	9,853,498	2008 Tweets	US
6	Boritz *et al.,* 2013	JA	ML	Dendograms in S+ (from HAC); Divisive Clustering; Concept Maps	Audit	387	2004-2009 SOX 404 Reports	Canada
7	Boukus & Rosenberg, 2006	WP	ML-LSA	LSA	Finance	152	FOMC Minutes	US
8	Brown & Tucker, 2011	JA	ML-TF-IDF	VSM; TF-IDF	Finance	28,142	1997-2006 MD&As	US
9	Buehlmaier, 2013	WP	ML-LR	LR	Finance	NA	News Articles	Hong Kong
10	Butler & Keselj, 2009	JA	ML-SVM	N-grams; CNG	Finance	NA	Annual Reports	Canada
11	Cecchini, 2005	DD	ML-SVM	SVM; Kernel methods	Audit	NA	AAERs Annual Reports	US
12	Cecchini *et al.,* 2010	JA	ML-SVM	SVM	Audit	NA	MD&As	US
13	Chakraborty, 2011	DD	ML-HAC	HAC Cluster analysis	Acctg	120		US

(*Continues*)

DOI: 10.1002/isaf

Appendix D: (Continued)

No.	Source	Lit Type	AI or ML	Method	Domain	N	Data Source	Origin
							Annual Reports Pension Notes	
14	Chakraborty & Vasarhelyi, 2010	CP	ML-HAC	HAC Cluster analysis	Acctg	120	Annual Reports	US
15	Chakraborty *et al.,* 2014	JA	ML-ANN	Decision Trees (DT) Naïve Bayes	Acctg	2,162	1984-2008 Journal Articles	US
16	Chen *et al.,* 2011	CP	ML-CRF	CRF	Audit	NA	Annual Reports	Taiwan
17	Chen *et al.,* 2013	JA	ML-CRF	CRF; MLR	Audit	22,780	Annual Reports	Taiwan
18	Cho *et al.,* 1998	JA	ML-SVM	Vector Weighting	Finance	392	Web News Articles	Hong Kong
19	Chung, 2014	JA	ML-ANN	BizPro ANN (Naïve Bayes)	Audit	231	News Articles	US
20	Core *et al.,* 2008	JA	ML	Tone-Measuring Algorithm	Finance	11,000	1994-2002 News Articles	US
21	Crofts & Bisman, 2010	JA	ML	Leximancer	Acctg	114	2000 - 2007 Acctg Journal Articles	US
22	Das & Chen, 2007	JA	ML	Classifier Algorithms Sentiment analysis	Finance	145,110	2001 Stock Message Boards	US
23	Engelberg, 2008	CP	ML	Typed Dependency Parsing (TDP)	Finance	80,935	Dow Jones News Articles	US
24	Fisher & Garnsey, 2006	JA	ML-LSI	LSI; CMU Toolkit; HAC	Acctg	567	FAS	US
25	Fisher & Garnsey, 2010	CP	ML	NA	Acctg	NA	FAS: Section 715	US
26	Fisher & McEwen, 2009	JA	ML	XML DTDs	Acctg	NA	FAS	US

(*Continues*)

DOI: 10.1002/isaf

Appendix D: (Continued)

No.	Source	Lit Type	AI or ML	Method	Domain	N	Data Source	Origin
27	Fung *et al.,* 2002	CP	ML-SVM	IA: IBM Intelligent Miner	Finance	350,000	News articles	Hong Kong
28	Fung *et al.,* 2005	JA	ML-SVM	SVM	Finance	350,000	News articles	Hong Kong
29	Fung *et al.,* 2003	CP	ML-SVM	SVM	Finance	614	News articles	Hong Kong
30	Gangolly & Tam, 2000	CP	ML	ML	Acctg	NA	2000 Annual Reports	US
31	Gangolly, 2008	CP	ML	NLP	Acctg	NA	FARS Database	US
32	Gangolly & Wu, 2000	JA	ML- Stats	Zipf's Law; TF-IDF	Acctg	232	FARS Database	US
33	Gangolly *et al.,* 1991	JA	AI-ES	ES	Acctg	NA	FAS	US
34	Garnsey, 2006a	JA	ML-LSI	LSI; HAC	Acctg	10	FAS	US
35	Garnsey, 2006b	CP	ML-LSI	LSI	Acctg	NA	FAS	US
36	Garnsey, 2008	CP	ML	NLP	Acctg	NA	FARS Database	US
37	Garnsey, 2009	CP	ML	NLP	Acctg	NA	FARS Database	US
38	Garnsey & Fisher, 2008	JA	ML-Stats	CMU Toolkit	Acctg	NA	FAS	US
39	Garnsey *et al.,* 2009	JA	ML-HAC	LSI & HAC	Acctg	127	FARS Database	US
40	Gerde, 2003	JA	ML	EDGAR-Analyzer; Intelligent Analysis (IA)	Acctg	18,595	Annual Reports	US
41	Gidofalvi, 2001	WP	ML-Stats	Rainbow Naïve Bayes	Finance	5,600	News Articles	US
42	Gilbert & Karahalios, 2010	CP	ML-Stats	Naïve Bayes	Finance	20,110,390	2008 LiveJournal Web Postings	US
43	Goel, 2008	DD	ML	Readability	Audit	NA	Annual Reports	US

(*Continues*)

DOI: 10.1002/isaf

Appendix D: (Continued)

No.	Source	Lit Type	AI or ML	Method	Domain	N	Data Source	Origin
44	Goel *et al.,* 2010	JA	ML-SVM	SVM; Naïve Bayes	Audit	126	Annual Reports	US
45	Grant & Conlon, 2006	JA	ML-KWIC	KWIC; CMU Toolkit	Acctg	283	Annual Reports	US
46	Groth & Muntermann, 2009	CP	ML-SVM	SVM	Finance	423	News articles	Germany
47	Groth & Muntermann, 2011	JA	ML-SVM	SVM; K-nn ANN: Naïve Bayes	Finance	423	EPRs	Germany
48	Gu *et al.,* 2007	JA	ML-ANN	Decision Tree	Finance	14	1998-2002 Web Postings	US
49	Hagenau *et al.,* 2013	JA	ML-SVM	SVM	Finance	14,348	1997-2011 EPRs; News articles	Germany
50	Hanley & Hoberg, 2010	JA	ML-SVM	Sentiment analysis Word vectors	Finance	2,043	IPO Prospectuses	US
51	Henry, 2006	JA	ML-LR	LR; CART OLS Regression	Finance	441	2002 EPRs	US
52	Hoberg & Phillips, 2009	JA	ML-SVM	Cosine Similarity	Finance	50,104	1997-2006 Annual Reports	US
53	Huang & Li, 2011	JA	ML-ANN	CKNN	Audit	21,077	2006-2010 Annual Reports	Singapore
54	Huang, Zang, & Zheng 2010	WP	ML-Stats	Naive Bayes	Finance	389,096	1995-2008 Analyst Reports	Hong Kong
55	Humpherys *et al.,* 2011	JA	ML-SVM	C4.5 Decision Tree; Naive Bayes; SVM	Audit	202	MD&As	US
56	Jiang, 2005	DD	ML	Semantic analysis	Finance	5,815	News Articles	US
57	Jin *et al.,* 2013	CP	ML-LDA	Sentiment analysis LDA; LMWL	Finance	361,782	2010 - 2013 News Articles	US

(*Continues*)

DOI: 10.1002/isaf

'Appendix D: (Continued)

No.	Source	Lit Type	AI or ML	Method	Domain	N	Data Source	Origin
58	Kamaruddin *et al.,* 2007	CP	ML	Conceptual Graphs	Audit		Annual Reports	Malaysia
59	Keila & Skillicorn, 2005	CP	ML-LIWC	LIWC; SVD	Audit	517,431	Enron Emails	Canada
60	Kumar *et al.,* 2012	JA	ML-SVM	Term Dictionary	Finance	149	News Articles	India
61	Larcker & Zakolyukina, 2012	JA	ML	LIWC	Finance	29,663	2003-2007 Transcripts of Conference Calls	US
62	Lavrenko *et al.,* 2000	CP	ML - Stats	Naïve Bayes Aenalyst	Finance	NA	40 Days of News Articles	US
63	Li, 2010a	CP	ML-Stats	Naive Bayes; LIWC	Finance	140,000	1994-2007 MD&As	US
64	Li, 2010b	JA	ML-Stats	Naïve Bayes	Finance	140,000	1994-2007 MD&As	US
65	Li, Wang, Li, Liu, Gong, & Chen 2014	JA	ML-SVM	SVR	Finance	124,470	Web Articles	Hong Kong
66	Li, Huang, Deng, & Zhu 2014	JA	ML-SVM	MKSVR	Finance	28,885	News Articles	Hong Kong
67	Li *et al.,* 2011	WP	ML-LIWC	Readability analysis FOG; LIWC	Finance	38,956	1994-2007 MD&As	US
68	Lu *et al.,* 2013	JA	MI-LR	LR	Finance	414	News Articles	Taiwan
69	Lugmayr & Gossen, 2012	CP	ML-SVM	SVM	Finance	NA	Broker Blog Postings	Finland
70	Magnusson *et al.,* 2005	JA	AI-SOM	SOM	Finance	24	Nokia, Ericsson, and Motorola 10Qs	Finland

(*Continues*)

DOI: 10.1002/isaf

Appendix D: (Continued)

No.	Source	Lit Type	AI or ML	Method	Domain	N	Data Source	Origin
71	Mahajan *et al.,* 2008	CP	ML-LDA	LDA; SVM Decision Tree	Finance	NA	Indian News Articles	India
72	Malo *et al.,* 2014	JA	ML-SVM	LPS; SVM	Finance	10,000	News Articles	Finland
73	Matsumoto *et al.,* 2011	JA	ML-LIWC	LIWC	Finance	10,062	Conference Calls	US
74	Mittermayer, 2004	CP	ML-SVM	SVM; TF-IDF	Finance	6,602	Press Releases	US
75	Mittermayer & Knolmayer, 2006	CP	ML-SVM	SVM; k-NN; TF-IDF	Finance	18,000	Press Releases	Switzerland
76	Mohan, 2006	WP	ML-LIWC	LIWC; GI	Finance	70,000	1996-2005 Annual Reports	US
77	Müller *et al.,* 1999	CP	ML-HAC	TaxGen	Acctg	70,000	One year of News articles	US
78	Peramunetilleke & Wong, 2002	JA	ML-Stats	TF-IDF	Finance	>400	News Articles	Australia
79	Purda & Skillicorn, 2014	JA	ML-SVM	Q-Tagger Random Forests; SVM	Audit	240	AAERs	Canada
80	Qiu *et al.,* 2013	JA	ML-SVM	TF-IDF; SVM	Audit	4,753	Chinese Annual Reports	China
81	Qiu *et al.,* 2014	JA	ML-SVM	SVM	Finance	5,421	1997-2003 Chinese Annual Reports	China
82	Rachlin *et al.,* 2007	CP	AI-DT	TF-IDF; GA; C4.5 DT	Finance	NA	News Articles	US
83	Schumaker & Chen, 2009	JA	ML-SVM	Named Entities AZFinText	Finance	9,211	News Articles	US
84	Schumaker *et al.,* 2012	JA	ML-SVM	AZFinText	Finance	2,802	News Articles	US

(*Continues*)

DOI: 10.1002/isaf

Appendix D: (Continued)

No.	Source	Lit Type	AI or ML	Method	Domain	N	Data Source	Origin
85	Seo *et al.,* 2002	WP	ML-LSA	LSA; SVD	Finance	1,239	News Articles	US
86	Sinha, 2010	WP	ML	Reuters Sentiment Engine Sentiment analysis	Finance	1,790,000	2003-2010 News Articles	US
87	Smales, 2014	JA	ML	MCQ Classifier Sentiment analysis	Finance	2,138,342	2000-2010 News Articles	Australia
88	Soni *et al.,* 2007	CP	ML	Concept Map	Finance	NA	News articles	India
89	Thomas & Sycara, 2002	JA	ML-Stats	Rainbow Naïve Bayes	Finance	200	News articles	US
90	Vu *et al.,* 2012	CP	ML-ANN	Decision-Tree; CRF TST; Content Maps	Finance	5,001,460	April - May 2011 Tweets	Vietnam
91	Wang & Wang, 2012	JA	ML-HAC	HAC	Finance	5	Annual Reports	China
92	Wang, Huang, & Wang 2012	JA	ML-CLAS	CLAS; POS	Finance	293	1995-2005 Annual Reports	China
93	Wuthrich *et al.,* 1998	CP	ML-SVM	k-NN	Finance	40	Hang Seng Index	Hong Kong
94	Yu *et al.,* 2013	JA	ML-Stats	Naïve Bayes	Finance	824	Social Media Postings	US
95	Zhai *et al.,* 2007	CP	ML-SVM	SVM; TF-IDF	Finance	148	2005-2006 News Articles	Australia

N: sample size; NA: not available or not applicable.

[a]Literature types: journal article (JA), conference paper (CP), working paper (WP), book chapter (BC) and doctoral dissertation (DD).

[b]Method: acronyms are defined in the glossary (Appendix A).

DOI: 10.1002/isaf

APPENDIX E: STUDIES EMPLOYING READABILITY ANALYSIS

Readability Studies & NLP							
No.	Source	Year	NLP Type	Method Used with Basic Readability Analysis	Domain	N	Origin
1	Baker and Kare 1992	1992	Manual		Acctg	NA	US
2	Brennan, Guillamon-Saorin, and Pierce 2009	2009	Manual		Acctg	NA	US
3	Butler and Keselj 2009	2009	ML-SVM	N-grams; CNG	Finance	NA	Canada
4	Clatworthy and Jones 2001	2001	Manual		Acctg	60	UK
5	Clatworthy and Jones 2003	2003	Manual		Acctg	100	UK
6	Clatworthy and Jones 2006	2006	Manual		Acctg	100	UK
7	Courtis 1986	1986	Manual		Acctg	142	Hong Kong
8	Courtis 1995a	1995	Manual		Acctg	32	Hong Kong
9	Courtis 1995b	1995	Manual		Acctg	NA	Hong Kong
10	Courtis 1998	1998	Text Mining		Audit	120	Hong Kong
11	De Franco, Hope, Vyas, and Zhou 2012	2012	Text Mining	FOG; Flesch	Finance	356,463	Canada
12	Dolphin and Wagley 1977	1977	Manual		Finance	NA	NA
13	Gibson and Schroeder 1990	1990	Manual		Acctg	40	US
14	Goel, et al. 2010	2010	ML-SVM	Naïve Bayes	Audit	126	US
15	Gu et al. 2007	2007	ML-ANN	Decision Tree	Finance	14	US
16	Healy 1977	1977	Manual		Acctg	NA	US
17	Henry 2006	2006	ML-LR	CART	Finance	441	US
18	Holley and Early 1980	1980	Manual		Acctg	NA	US
19	Humpherys et al. 2011	2011	ML-SVM	C4.5 Decision Tree; Naive Bayes	Audit	202	US
20	Jones 1988	1988	Manual		Acctg	32	UK

(Continues)

DOI: 10.1002/isaf

Appendix E: (Continued)

21	Kravet and Muslu 2011	2011	Text Mining	Parsing	Finance	4,315	US
22	Lawrence 2011	2011	Manual		Finance	91,228	US
23	Lebar 1982	1982	Manual		Acctg	10	US
24	Lee 2010	2010	Text Mining	FOG	Finance	60,161	Taiwan
25	Lehavy, Li, and Merkley 2011	2011	Manual		Finance	33,704	US
26	Li 2008	2008	Text Mining	FOG	Acctg	55,719	US
27	Li, Lundholm, and Minnis 2011	2011	ML-LIWC	FOG	Finance	38,956	US
28	Loughran and McDonald 2009	2009	Text Mining	FOG; Flesch	Finance	42,357	US
29	Loughran and McDonald 2010	2010	Text Mining	FOG; Flesch; Plain English	Finance	42,357	US
30	Loughran and McDonald 2014	2014	Text Mining	FOG	Finance	66,707	US
31	Loughran, McDonald, and Yun 2009	2009	Manual		Audit	67,044	US
32	Merkley 2014	2014	Text Mining	Sentiment analysis; FOG	Finance	22,482	US
33	Miller 2010	2010	Text Mining	FOG	Finance	12,771	US
34	Moffitt and Burns 2009	2009	Text Mining	Content analysis; NLTK	Audit	720	US
35	Nelson and Pritchard 2007	2007	Text Mining	FOG	Acctg	53,315	US
36	Othman et al. 2012	2012	Text Mining	Flesch	Finance	16	Malaysia
37	Pashalian and Crissy 1952	1952	Manual		Acctg	NA	NA
38	Qiu, Jiang, and Deng 2013	2013	ML-SVM	TF-IDF	Audit	4,753	China
39	Rees and Twedt 2012	2012	Text Mining	Sentiment analysis; GI	Finance	2,057	US
40	Schroeder and Gibson 1990	1990	Manual		Acctg	40	US
41	Smith and Smith 1971	1971	Manual		Acctg	50	US
42	Smith and Taffler 1992a	1992	Manual		Acctg	66	US
43	Smith and Taffler 1992b	1992	Manual		Acctg	NA	Australia
44	Soper and Dolphion 1964	1964	Manual		Acctg	25	US
45	Still 1972	1972	Manual		Finance	NA	US
46	Sydserff and Weetman 1999	1999	Manual		Audit	10	UK
47	Twedt and Rees 2012	2012	Text Mining	Sentiment analysis; GI; FOG	Finance	2,057	US

N: sample size; NA: not available or not applicable.

[a]Method: acronyms are defined in the glossary (Appendix A).

DOI: 10.1002/isaf

APPENDIX F: NLP STUDIES CATEGORIZED BY ASSOCIATED OR CO-OCCURRING METHOD

	AI	ML	ANN	CART	CMU	Concept Maps & Graphs	CRF	DT	FOG	HAC	IA	k-NN
Classification Studies												
KO, Categorization, Information Retrieval	1	5			3				1	1	1	
Taxonomy/Thesauri Generation		13	1		3	1				6		
Information Retrieval		5	1			1		1		2		
Assessing F/S Content	1	4				1	2					
Prediction Studies												
Fraud Prediction and Dection		8						1				
Stock Prices and Market Activity	2	32	3			1		1			1	3
Predictive Value of Annual Reports/Disclosures		11		1								
Predictive Value of Web Content		7	3				1	2				
Total Excluding Readability	4	85	8	1	6	4	3	5	1	9	2	3
NLP-Based Readability Studies		6						1	4			
Total All Groups	4	91	8	1	6	4	3	6	5	9	2	3

(Continued)

DOI: 10.1002/isaf

Appendix F: (Continued)

	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC
Classification Studies										
KO, Categorization, Information Retrieval				1						
Taxonomy/Thesauri Generation				1	1					
Information Retrieval				1		1				
Assessing F/S Content					1					
Prediction Studies										
Fraud Prediction and Dection		2		1		2				1
Stock Prices and Market Activity	2	3	1	3		5	2	1	4	
Predictive Value of Annual Reports/Disclosures	1	1	1			2		2	1	
Predictive Value of Web Content						3	1		1	
Total Excluding Readability	3	6	2	7	2	13	3	3	6	1
NLP-Based Readability Studies			1			1		1		7
Total All Groups	3	6	3	7	2	14	3	4	6	8

(Continued)

DOI: 10.1002/isaf

Appendix F: (Continued)

	AI	SA	SOM	Stats	SVM	TF IDF	VSM	Other*
Classification Studies								
KO, Categorization, Information Retrieval	1							1
Taxonomy/Thesauri Generation				2	1	2		3
Information Retrieval								2
Assessing F/S Content	1		1		1	1		
Prediction Studies								
Fraud Prediction and Dection						6		
Stock Prices and Market Activity	2	5	1	4	19	5	3	5
Predictive Value of Annual Reports/Disclosures		1		3	3		2	1
Predictive Value of Web Content		3		2	1			2
Total Excluding Readability	4	9	2	11	31	8	5	14
NLP-Based Readability Studies		1			2	1		1
Total All Groups	4	10	2	11	33	9	5	15

*Method was only used in one study.

DOI: 10.1002/isaf

APPENDIX G: SUMMARY OF ALL 262 STUDIES CATEGORIZED BY DATA SOURCE AND ANALYSIS METHOD

Study	Data Source	Manual	Text Mining	ML	AI	ANN	CART	CMU	Concept Maps/ Graphs	CRF	DT	FOG	HAC	IA	k-NN	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC	SA	SOM	Stats	SVM	TF-IDF	VSM	Other
Back et al. 2001	Annual Reports & CEO Reports				X	X																					X					
Baker & Kare 1992	Annual Reports	X																														
Balakrishnan Qiu & Srinivasan 2010	Annual Reports			X																									X			
Bao & Datta 2014	Annual Reports			X												X																
Boo & Simnett 2002	Annual Reports	X																														
Bovee et al. 2005	Annual Reports		X																													
Bowman 1984	Annual Reports	X																														
Butler & Keselj 2009	Annual Reports			X																									X			X
Campbell et al. 2014	Annual Reports		X																													
Cecchini 2005	Annual Reports			X																									X			
Chakraborty & Vasarhelyi 2010	Annual Reports			X									X																			
Chakraborty 2011	Annual Reports			X									X																			
Chen et al. 2011	Annual Reports			X						X																						
Chen et al. 2013	Annual Reports			X						X										X												
Clapham & Schwenk 1991	Annual Reports	X																														
Clarkson Kao & Richardson 1994	Annual Reports	X																														
Clarkson Kao & Richardson 1999	Annual Reports	X																														
Clatworthy & Jones 2001	Annual Reports	X																														
Clatworthy & Jones 2003	Annual Reports	X																														
Clatworthy & Jones 2006	Annual Reports	X																														
Cong Kogan & Vasarhelyi 2007	Annual Reports		X																													
Courtis 1986	Annual Reports	X																														
Courtis 1995a	Annual Reports	X																														
Courtis 1995b	Annual Reports	X																														
Craig Mortensen & Iyer 2013	Annual Reports		X																													
Dolphin & Wagley 1977	Annual Reports	X																														
Ferguson 1997	Annual Reports		X																													
Folsom et al. 2013	Annual Reports	X																														
Frazier Ingram & Tennyson 1984	Annual Reports		X																													
Gangolly & Tam 2000	Annual Reports			X				X					X																	X		
Gerdes 2003	Annual Reports			X										X																		
Gibson & Schroeder 1990	Annual Reports	X																														
Goel 2008	Annual Reports			X																					X				X			
Goel et al. 2010	Annual Reports			X																	X								X			
Grant & Conlon 2006	Annual Reports			X				X																								X
Healy 1977	Annual Reports	X																														
Hoberg & Phillips 2009	Annual Reports			X																			X						X		X	
Holley & Early 1980	Annual Reports	X																														
Huang & Li 2011	Annual Reports			X		X																										X

(Continues)

DOI: 10.1002/isaf

Appendix G: (Continued)

Study	Data Source	Manual	Text Mining	ML	AI	ANN	CART	CMU	Concept Maps/ Graphs	CRF	DT	FOG	HAC	IA	k-NN	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC	SA	SOM	Stats	SVM	TF-IDF	VSM	Other
Hussainey Schleicher & Walker 2003	Annual Reports		X																													
Ingram & Frazier 1980	Annual Reports	X																														
Jegadeesh & Wu 2013	Annual Reports		X																													
Jones 1988	Annual Reports	X																														
Kamaruddin Hamdan and Bakar 2007	Annual Reports			X					X																							
Klamm & Watson 2009	Annual Reports	X																														
Kothari Li & Short 2009	Annual Reports & 10-Qs	X																														
Kravet & Muslu 2011	Annual Reports		X																													
Lawrence 2011	Annual Reports	X																														
Lebar 1982	Annual Reports & EPRs	X																														
Lehavy Li & Merkley 2011	Annual Reports	X																														
Leinemann et al. 2001	Annual Reports		X																													
Levine & Smith 2006	Annual Reports		X																													
Li 2006	Annual Reports	X																														
Li 2008	Annual Reports		X																													
Linsley & Shrives 2006	Annual Reports	X																														
Loughran & McDonald 2009	Annual Reports		X																													
Loughran & McDonald 2010	Annual Reports		X									X													X							
Loughran & McDonald 2011a	Annual Reports		X																													
Loughran & McDonald 2011b	Annual Reports		X																													
Loughran & McDonald 2014	Annual Reports		X									X													X							
Loughran McDonald & Yun 2009	Annual Reports	X																														
Merkley 2011	Annual Reports		X																													
Miller 2010	Annual Reports		X									X													X							
Mirakur 2011	Annual Reports		X																													
Moffit & Burns 2011	Annual Reports		X																													
Mohan 2006	Annual Reports			X													X															X
Muslu et al. 2010	Annual Reports		X																													
Nelson & Pritchard 2007	Annual Reports		X																													
Pashalian & Crissy 1952	Annual Reports	X																														
Qiu Jiang & Deng 2013	Annual Reports			X																									X	X		
Qiu Srinivasan and Hu 2014	Annual Reports			X																									X			
Rutherford 2005	Annual Reports	X																														
Singhvi 1968	Annual Reports	X																														
Smith & Smith 1971	Annual Reports	X																														
Soper & Dolphion 1964	Annual Reports	X																														
Steele 1982	Annual Reports	X																														
Summers & Sweeney 1998	Annual Reports	X																														
Tennyson Ingram & Dugan 1990	Annual Reports		X																													
Wang & Wang 2012	Annual Reports			X									X																			
Wang Li & Cao 2012	Annual Reports		X																													
You & Zhang 2009	Annual Reports	X																														

(Continues)

DOI: 10.1002/isaf

Appendix G: (Continued)

Study	Data Source	Manual	Text Mining	ML	AI	ANN	CART	CMU	Concept Maps/ Graphs	CRF	DT	FOG	HAC	IA	k-NN	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC	SA	SOM	Stats	SVM	TF-IDF	VSM	Other
Antweiler & Frank 2006	News Articles		X																													
Buehlmaier 2013	News Articles			X														X														
Carretta et al. 2011	News Articles		X																													
Chakraborty Chiu & Vasarhelyi 2014	News Articles			X		X					X										X											
Chung 2014	News Articles			X		X															X											
Cohen et al. 2010	News Articles	X																														
Core Guay & Larcker 2008	News Articles			X																			X								X	
Engelberg 2008	News Articles			X																												X
Engelberg Reed & Ringgenberg 2012	News Articles		X																													
Ferguson et al. 2014	News Articles		X																													
Fung Yu & Lam 2002	News Articles			X										X															X			
Fung Yu & Lu 2005	News Articles			X																									X			
Fung Yu & Wai 2003	News Articles			X																									X			
Garcia 2014	News Articles	X																														
Gidofalvi 2001	News Articles			X																	X							X				
Groth & Muntermann 2009	News Articles			X																									X			
Huang et al. 2010	News Articles		X																													
Jacobs & Rau 1990	News Articles		X																													
Jiang 2005	News Articles			X																		X		X		X						
Jin et al. 2013	News Articles			X												X										X						X
Kumar Kumar & Prasad 2012	News Articles			X																									X			
Lavrenko et al. 2000	News Articles			X																	X							X				
Li Huang Deng & Zhu 2014	News Articles			X																									X			
Lu Shen & Wei 2013	News Articles			X														X														
Mahajan Dey & Haque 2008	News Articles			X							X					X													X			
Malo et al. 2014	News Articles			X																									X			X
Muller et al. 1999	News Articles			X									X																			
Peramunetilleke & Wong 2002	News Articles			X																								X		X		
Rachlin et al. 2007	News Articles				X						X																			X		X
Schumaker & Chen 2009	News Articles			X																		X							X			
Schumaker et al. 2012	News Articles			X																									X			
Seo Giampaca & Sycara 2002	News Articles			X															X													
Sinha 2010	News Articles			X																						X						
Smales 2014	News Articles			X																						X						
Soni Van Eck & Kaymack 2007	News Articles			X					X																				X			
Tetlock 2007	News Articles		X																													
Tetlock 2011	News Articles		X																													
Tetlock Saar-Tsechansky & Macskassy 200	News Articles		X																													
Thomas & Sycara 2002	News Articles			X																	X							X				
Zhai Hsu & Halgamuge 2007	News Articles			X																				X					X	X		
Brown & Tucker 2011	MD&AS			X																										X	X	

(Continues)

DOI: 10.1002/isaf

Appendix G: (Continued)

Study	Data Source	Manual	Text Mining	ML	AI	ANN	CART	CMU	Concept Maps/ Graphs	CRF	DT	FOG	HAC	IA	k-NN	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC	SA	SOM	Stats	SVM	TF-IDF	VSM	Other
Bryan 1997	MD&As	X																														
Burns et al. 2010	MD&As		X																													
Callahan & Smith 2004	MD&As	X																														
Cecchini et al. 2010	MD&As			X																								X				
Cole 1990	MD&As	X																														
Feldman et al. 2008	MD&As		X																						X							
Feldman et al. 2010	MD&As		X																													
Hooks & Moon 1993	MD&As	X																														
Humpherys et al. 2011	MD&As			X							X																	X				
Li 2010a	MD&As			X													X				X			X			X					
Li 2010b	MD&As			X																	X			X			X					
Li Lundholm & Minnis 2011	MD&As			X								X					X								X							
Moffitt & Burns 2009	MD&As		X																													
Pava & Epstein 1993	MD&As	X																														
Schroeder & Gibson 1990	MD&As	X																														
Sydserff & Weetman 1999	MD&As	X																							X							
Wang Huang & Wang 2012	MD&As			X																			X	X								X
Andersen et al. 1992	EPRs		X																													
Davis & Tama-Sweet 2012	EPRs		X																													
Davis Piger & Sedor 2006	EPRs		X																													
Davis Piger & Sedor 2012	EPRs		X																													
Francis Schipper & Vincent 2002	EPRs	X																														
Groth & Muntermann 2011	EPRs			X		X									X						X							X				
Hagenau Liebman & Neumann 2013	EPRs + News articles			X																								X				
Henry & Leone 2010	EPRs		X																													
Henry 2006	EPRs			X			X											X					X		X							
Hoskin Hughes & Ricks 1986	EPRs	X																														
Huang 2011	EPRs	X																														
Huang Teoh & Zhang 2014	EPRs	X																														
Kimbrough & Wang 2014	EPRs	X																														
Rogers Van Buskirk & Zechman 2011	EPRs		X																													
Fisher & Garnsey 2006	FAS			X				X					X						X													
Fisher & Garnsey 2010	FAS			X																												
Fisher 2004	FAS		X																													
Fisher 2007	FAS		X																													
Fisher and McEwen 2009	FAS			X																												X
Gangolly Hedley & Wong 1991	FAS				X																											X
Garnsey & Fisher 2008	FAS			X				X												X								X				
Garnsey 2006a	FAS			X				X					X						X													
Garnsey 2006b	FAS			X				X											X													

(Continues)

DOI: 10.1002/isaf

Appendix G: (Continued)

Study	Data Source	Manual	Text Mining	ML	AI	ANN	CART	CMU	Concept Maps/ Graphs	CRF	DT	FOG	HAC	IA	k-NN	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC	SA	SOM	Stats	SVM	TF-IDF	VSM	Other
Gangolly & Wu 2000	FARS Database			X																								X		X		X
Gangolly 1989	FARS Database		X																													
Gangolly 2008	FARS Database			X																												X
Garnsey & Hotaling 2011	FARS Database		X																													
Garnsey 2008	FARS Database			X																												
Garnsey 2009	FARS Database			X																												
Garnsey O'Neil & Stokes 2009	FARS Database			X									X						X													
Beynon Clatworthy & Jones 2005	Chairman's Statements		X																													
Courtis 1998	Chairman's Statements		X																													
Othman et al. 2012	Chairman's Statements		X																													
Smith & Taffler 1992a	Chairman's Statements	X																														
Smith & Taffler 1992b	Chairman's Statements	X																														
Smith & Taffler 2000	Chairman's Statements	X																														
Still 1972	Chairman's Statements	X																														
Sydserff & Weetman 2002	Chairmans' statements		X																													
Bricker et al. 1995	Analyst Reports	X																														
De Franco Hope Vyas and Zhou 2012	Analyst Reports		X																													
Govindarajan 1980	Analyst Reports	X																														
Huang Zang & Zheng 2010	Analyst Reports			X																	X							X				
Previts et al. 1994	Analyst Reports	X																														
Rees & Twedt 2012	Analyst Reports		X																													
Rogers & Grant 1997	Analyst Reports & Annual Reports	X																														
Twedt & Rees 2012	Analyst Reports		X																													
Brennan Guillamon-Saorin Pierce 2009	Press Releases	X																														
Henry 2008	Press Releases		X																													
Mittermayer & Knolmayer 2006	Press Releases			X											X														X	X		
Mittermayer 2004	Press Releases			X																									X	X		
Ozik & Sadka 2012	Press Releases		X																													
Sadique In & Veeraraghavan 2008	Press releases		X																													
Solomon Soltes & Sosyuram 2014	Press releases		X																													
Demers & Vega 2010	10-Qs		X																													
Lee 2010	10-Qs		X																													
Magnusson et al 2005	10Qs				X																						X					
Miller 2002	10-Qs	X																														
Shon 2003	10-Qs	X																														
Deumes 2008	IPO Prosoectuses	X																														
Ferris Hao & Liao 2013	IPO Prospectuses		X																													
Hanley & Hoberg 2008	IPO Prospectuses		X																													
Hanley & Hoberg 2010	IPO Prospectuses			X																						X			X		X	
Jarvenpaa & Ives 1990	Letters to Shareholders	X																														

(Continues)

DOI: 10.1002/isaf

Appendix G: (Continued)

Study	Data Source	Manual	Text Mining	ML	AI	ANN	CART	CMU	Concept Maps/ Graphs	CRF	DT	FOG	HAC	IA	k-NN	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC	SA	SOM	Stats	SVM	TF-IDF	VSM	Other
Abrahamson & Amir 1996	President's Letters	X																														
Abrahamson & Park 1994	President's Letters	X																														
McConnell Haslem & Gibson 1986	Presidents' Letters	X																														
Swales 1988	President's Letters	X																														
Churyk Lee & Clinton 2009	AAERs	X																														
Dechow et al. 2009	AAERs	X																														
Purda & Skillicorn 2014	AAERs			X																								X				
Larcker & Zakolyukina 2012	Conference Calls			X													X															
Matsumoto Pronk & Roelofsen 2011	Conference Calls			X													X															
Price et al. 2012	Conference Calls		X																													
Aerts 1994	Disclosures	X																														
Beattie McInnes & Fearnley 2004	Disclosures		X																													
Bhojraj Blacconiere & D'Souza 2004	Disclosures	X																														
Bollen Mao & Pepe 2011	Tweets		X																													X
Bollen Mao & Zeng 2011	Tweets			X		X																										
Vu Chang Ha and Collier 2012	Tweets			X		X			X	X	X											X		X		X						X
Crofts & Bisman 2010	Acctg Journal Articles			X																												
Previts & Brown 1993	Acctg Journal Articles	X																														
Armesto et al. 2009	Beige Book		X																													
Sadique et al. 2013	Beige Book		X																													
Lugmayr & Gossen 2012	Blog Postings			X																								X				
Luo Zhang & Duan 2013	Blog Postings	X																														
D'Aveni & MacMillan 1990	CEO Letters to Shareholders	X																														
Hooghiemstra 2010	CEO Letters to Shareholders	X																														
Davis Ge Matsumoto & Zhang 2014	Earnings Calls		X																													
Doran Peterson & Price 2010	Earnings Calls		X																													
Kasznik & Lev 1995	Mgmt. Disclosures	X																														
Kelly-Newton 1980	Mgmt. Disclosures	X																														
Boritz Hayes & Lim 2013	SOX 404 Reports			X					X				X																			
Masli et al. 2009	SOX 404 Reports	X																														
Hofstedt 1976	Various	X																														
Ijiri Kinard and Putney 1968	Various	X																														
Cho Wuthrich & Zhang 1998	Web News Articles			X																								X		X		
Li Wang Li Liu Gong & Chen 2014	Web News Articles			X																								X				
Gilbert & Karahalios 2010	Web postings			X																	X							X				
Gu et al. 2007	Web postings			X		X					X																					
Badua 2005	Accounting Research Database	X																														
Rezaee Olibe & Minmier 2003	Audit Committee Reports	X																														
Meyer & Rigsby 2001	BRIA articles	X																														
Farrell & Cobbin 2000	Codes of Ethics	X																														
Mendonca 2009	CRSP		X																													
Keila & Skillicorn 2005	Emails			X													X		X													

(Continues)

DOI: 10.1002/isaf

Appendix G: (Continued)

Study	Data Source	Manual	Text Mining	ML	AI	ANN	CART	CMU	Concept Maps/ Graphs	CRF	DT	FOG	HAC	IA	k-NN	LDA	LIWC	LR	LSA LSI SVD	MLR	NB	NER	OLS	POS	RBC	SA	SOM	Stats	SVM	TF-IDF	VSM	Other
Yen Hirst & Hopkins 2007	FASB Comment Letters	X																														
Boukus & Rosenberg 2006	FOMC Minutes			X															X													
Born Ehrmann & Fratzscher 2013	FSRs & Speeches		X																													
Bligh & Hess 2007	Greenspan Speeches		X																													
Wuthrich et al. 1998	Hang Seng Index (HIS)			X											X														X			
Antweiler & Frank 2004	Message Boards			X																	X								X			
Baginski Hassell & Kimbrough 2004	Mgmt earning forecasts	X																														
Chen et al. 2014	Online Postings		X																													
Wang & Guo 2012	Online Recruiting Information		X																													
Ding & Chen 2006	Proxy Statements		X																													
Merkley 2014	R&Ds		X									X													X	X						
Durnev & Mangen 2011	Restatements		X																													
Loughran & McDonald 2013b	S-1 Filings		X																													
Donelson McInnis & Mergenthaler 2012	Securities Lawsuits	X																														
Yu Duan & Cao 2013	Social Media Postings			X																	X					X		X				
Das & Chen 2007	Stock message boards			X																						X						
Chiu 2013	TAR articles		X																													
Lytinen & Gershman 1986	Telex messages		X																													
Total Studies per Method		**86**	**81**	**91**	**4**	**8**	**1**	**6**	**4**	**3**	**6**	**5**	**9**	**2**	**3**	**3**	**6**	**3**	**7**	**2**	**14**	**3**	**4**	**6**	**8**	**10**	**2**	**11**	**33**	**9**	**5**	**15**

REFERENCES

Abrahamson E, Amir E. 1996. The information content of the president's letter to shareholders. *Journal of Business Finance & Accounting* **23**(8): 1157–1182.

Abrahamson E, Park C. 1994. Concealment of negative organizational outcomes: an agency theory perspective. *Academy of Management Journal* **37**(5): 1302–1334.

Aerts W. 1994. On the use of accounting logic as an explanatory category in narrative accounting disclosures. *Accounting, Organizations and Society* **19**(4–5): 337–353.

Andersen PM, Hayes PJ, Huettner AK, Schmandt LM, Nirenburg IB, Weinstein SP. 1992. Automatic extraction of facts from press releases to generate news stories. Paper read at *Third Conference on Applied Natural Language Processing*.

Antweiler W, Frank MZ. 2004. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance* **59**(3): 1259–1294.

Antweiler W, Frank M. 2006. Do U.S. stock markets typically overreact to corporate news stories? Working paper, University of British Columbia.

Armesto MT, Hernández-Murillo R, Owyang MT, Piger JM. 2009. Measuring the information content of the Beige Book: a mixed data sampling approach. *Journal of Money, Credit and Banking* **41**(1): 35–55.

Back B, Toivonen J, Vanharanta H, Visa A. 2001. Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems* **2**(4): 249–269.

Badua FA. 2005. Pondering paradigms: Tracing the development of accounting thought with taxonomic and citation analysis. Doctoral dissertation, Rutgers The State University of New Jersey, ProQuest Dissertations & Theses Database.

Baginski S, Hassell J, Kimbrough M. 2004. Why do managers explain their earnings forecasts? *Journal of Accounting Research* **42**(1): 1–29.

Baker HEI, Kare D. 1992. Relationship between annual report readability and corporate financial performance. *Management Research News* **15**(1): 1–4.

Balakrishnan R, Qiu XY, Srinivasan P. 2010. On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research* **202**(3): 789–801.

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

Beattie V, McInnes W, Fearnley S. 2004. A methodology for analysing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. *Accounting Forum* **28**(3): 205–236.

Beynon MJ, Clatworthy MA, Jones MJ. 2005. The prediction of profitability using accounting narratives: a variable-precision rough set approach. *Intelligent Systems in Accounting and Finance Management* **12**(4): 227–242.

DOI: 10.1002/isaf

Bhojraj S, Blacconiere WG, D'Souza JD. 2004. Voluntary disclosure in a multi-audience setting: an empirical investigation. *The Accounting Review* **79**(4): 921–947.
Bligh M, Hess GD. 2007. The power of leading subtly: Alan Greenspan, rhetorical leadership, and monetary policy. *Leadership Quarterly* **18**(2): 87–104.
Bollen J, Mao H, Pepe A. 2011a. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Paper read at *Fifth International AAAI Conference on Weblogs and Social Media*.
Bollen J, Mao H, Zeng XJ. 2011b. Twitter mood predicts the stock market. *Journal of Computational Science* **2**(1): 1–8.
Boo E, Simnett R. 2002. The information content of management's prospective comments in financially distressed companies: a note. *Abacus* **38**(2): 280–295.
Boritz JE, Hayes L, Lim J-H. 2013. A content analysis of auditors' reports on IT internal control weaknesses: the comparative advantages of an automated approach to control weakness identification. *International Journal of Accounting Information Systems Research* **14**(2): 138–163.
Born B, Ehrmann M, Fratzscher M. 2013. Central bank communication on financial stability. *The Economic Journal* **124**(577): 701–734.
Boukus E, Rosenberg JV. 2006. The information content of FOMC minutes. Federal Reserve Bank of New York: New York.
Bovee M, Kogan A, Nelson K, Srivastava RP, Vasarhelyi MA. 2005. Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL). *Journal of Information Systems* **19**(1): 19–41.
Bowman EH. 1984. Content analysis of annual reports for corporate strategy and risk. *Interfaces* **14**(1): 61–71.
Brennan NM, Guillamon-Saorin E, Pierce A. 2009. Impression management: developing and illustrating a scheme of analysis for narrative disclosures—a methodological note. *Accounting, Auditing & Accountability Journal* **22**(5): 789–832.
Bricker R, Previts G, Robinson T, Young S. 1995. Financial analyst assessment of company earnings quality. *Journal of Accounting, Auditing & Finance* **10**(3): 541–555.
Brown C, Wong J, Baldwin A. 2007. A review and analysis of existing streams of research in continuous auditing. *Journal of Emerging Technologies in Accounting* **4**(1): 1–28.
Brown S, Tucker J. 2011. Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research* **49**(2): 309–346.
Bryan SH. 1997. Incremental information content of required disclosures contained in management discussion and analysis. *The Accounting Review* **72**(2): 285–301.
Buehlmaier MMM. 2013. The role of the media in takeovers: theory and evidence. Working paper, The University of Hong Kong.
Burke JF. 1997. Report on standards overload. *The CPA Journal* **66**(3): 11.
Burns M, Moffitt K, Felix W, Burgoon J. 2010. Using lexical bundles to discriminate between fraudulent and non-fraudulent financial reports. University of Arizona.
Butler M, Keselj V. 2009. Financial forecasting using character *N*-gram analysis and readability scores of annual reports. In Advances in Artificial Intelligence, Gao Y, Japkowicz N (eds). Springer: Berlin; 39–51.
Callahan CM, Smith R. 2004. Firm performance and management's discussion and analysis disclosures: an industry approach. Paper read at University of Arkansas, Oklahoma State University, and Central States Research Conference.
Campbell JL, Chen HC, Dhaliwal DS, Lu HM, Steele LB. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies* **19**(1): 396–455.
Carretta A, Farina V, Graziano EA, Reale M. 2011. Does investor attention influence stock market activity? The case of spin-off deals. MPRA Paper No. 33545.
Cecchini M. 2005. Quantifying the risk of financial events using kernel methods and information retrieval. University of Florida.
Cecchini M, Aytug H, Koehler GJ, Pathak P. 2010. Making words work: using financial text as a predictor of financial events. *Decision Support Systems* **50**(1): 164–175.
Chakraborty V. 2011. Three essays on using text analytic techniques for accounting research. Doctoral dissertation, Rutgers, The State University of New Jersey, ProQuest Dissertations & Theses Database.
Chakraborty V, Vasarhelyi M. 2010. Automating the process of taxonomy creation and comparison of taxonomy structures. Paper read at *Collected Papers of the 19th Annual Research Workshop on Strategic and Emerging Technologies*, American Accounting Association.

DOI: 10.1002/isaf

Chakraborty V, Chiu V, Vasarhelyi M. 2014. Automatic classification of accounting literature. *International Journal of Accounting Information Systems* **15**(2): 122–148.
Chen CL, Liu CL, Chang YC, Tsai HP. 2011. Exploring the relationships between annual earnings and subjective expressions in US financial statements. Paper read at *8th International Conference on e-Business Engineering* (*ICEBE*), Beijing, China.
Chen CL, Liu CL, Chang YC, Tsai HP. 2013. Opinion mining for relating subjective expressions and annual earnings in US financial statements. *Journal of Information Science and Engineering* **29**(4): 743–764.
Chen H, De P, Hu YJ, Hwang BH. 2014. Wisdom of crowds: the value of stock opinions transmitted through social media. *Review of Financial Studies* **27**(5): 1367–1403.
Chiu V. 2013. Accounting bibliometrics: the development and intellectual structure of accounting research. Rutgers, The State University of New Jersey: Newark, NJ.
Cho V, Wutrich B, Zhang J. 1998. Text processing for classification. *Journal of Computational Intelligence in Finance* **7**(2): 6–22.
Chung W. 2014. BizPro: extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management* **34**(2): 272–284.
Churyk NT, Lee C-C, Clinton BD. 2009. Early detection of fraud: evidence from restatements. In Advances in Accounting Behavioral Research, Arnold V (ed.). Emerald Group Publishing Limited; 25–40.
Clapham SE, Schwenk CR. 1991. Self-serving attributions, managerial cognition, and company performance. *Strategic Management Journal* **12**(3): 219–229.
Clarkson PM, Kao JL, Richardson GD. 1994. The voluntary inclusion of forecasts in the MD&A section of annual reports. *Contemporary Accounting Research* **11**(1): 423–450.
Clarkson PM, Kao JL, Richardson GD. 1999. Evidence that management discussion and analysis (MD&A) is a part of a firm's overall disclosure package. *Contemporary Accounting Research* **16**(1): 11–34.
Clatworthy M, Jones MJ. 2001. The effect of thematic structure on the variability of annual report readability. *Accounting, Auditing & Accountability Journal* **14**(3): 311–326.
Clatworthy M, Jones MJ. 2003. Financial reporting of good news and bad news: evidence from accounting narratives. *Accounting and Business Research* **33**(3): 171–185.
Clatworthy M, Jones MJ. 2006. Differential patterns of textual characteristics and company performance in the chairman's statement. *Accounting, Auditing & Accountability Journal* **19**(4): 493–511.
Cohen JR, Ding Y, Lesage C, Stolowy H. 2010. Corporate fraud and managers' behavior: evidence from the press. *Journal of Business Ethics* **95**(s2): 271–315.
Cole C. 1990. MD&A trends in Standard & Poor's top 100 companies. *Journal of Corporate Accounting & Finance* **2**(2): 127–136.
Cong Y, Kogan A, Vasarhelyi MA. 2007. Extraction of structure and content from the EDGAR database: a template-based approach. *Journal of Emerging Technologies in Accounting* **4**(1): 69–86.
Core JE, Guay W, Larcker DF. 2008. The power of the pen and executive compensation. *Journal of Financial Economics* **88**(1): 1–25.
Courtis JK. 1986. An investigation into annual report readability and corporate risk return relationships. *Accounting and Business Research* **16**(64): 285–294.
Courtis JK. 1995a. Readability of the annual report: Western versus Asian evidence. *Accounting, Auditing & Accountability Journal* **8**(2): 4–17.
Courtis JK. 1995b. Readability of financial statements. *The Hong Kong Accountant* **6**(4): 66–74.
Courtis JK. 1998. Annual report readability variability: tests of the obfuscation hypothesis. *Accounting, Auditing & Accountability Journal* **11**(4): 459–472.
Craig R, Mortensen T, Iyer S. 2013. Exploring top management language for signals of possible deception: the words of Satyam's chair Ramalinga Raju. *Journal of Business Ethics* **113**(2): 333–347.
Crofts K, Bisman J. 2010. Interrogating accountability: an illustration of the use of Leximancer software for qualitative data analysis. *Qualitative Research in Accounting & Management* **7**(2): 180–207.
D'Aveni RA, MacMillan IC. 1990. Crisis and the content of managerial communications: a study of the focus of attention of top managers in surviving and failing firms. *Administrative Science Quarterly* **35**(4): 634–657.
Das SR, Chen MY. 2007. Yahoo! for Amazon: sentiment extraction from small talk on the web. *Management Science* **53**(9): 1375–1388.
Davis AK, Tama-Sweet I. 2012. Managers' use of language across alternative disclosure outlets: earnings press releases versus MD&A. *Contemporary Accounting Research* **29**(3): 804–837.

DOI: 10.1002/isaf

Davis AK, Piger JM, Sedor LM. 2006. Beyond the numbers: an analysis of optimistic and pessimistic language in earnings press releases. Federal Reserve Bank of St. Louis Working Paper Series.
Davis AK, Piger JM, Sedor LM. 2012. Beyond the numbers: measuring the information content of earnings press release language. *Contemporary Accounting Research* **29**(3): 845–868.
Davis AK, Ge W, Matsumoto D, Zhang JL. 2014. The effect of manager-specific optimism on the tone of earnings conference calls. Paper read at *CAAA Annual Conference 2012*.
De Franco G, Hope O, Vyas D, Zhou Y. 2012. Analyst report readability. *Contemporary Accounting Research* **32**(1): 76–104.
Dechow PM, Ge W, Larson CR, Sloan RG. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* **28**(1): 17–82.
Demers EA, Vega C. 2010. Soft information in earnings announcements: news or noise? INSEAD & Board of Governors of the Federal Reserve System: Washington, DC.
Deumes R. 2008. Corporate risk reporting: a content analysis of narrative risk disclosures in prospectuses. *Journal of Business Communication* **45**(2): 120–157.
Ding C, Chen P. 2006. Mining executive compensation data from SEC Filings Paper read at *22nd International Conference on Data Engineering Workshops*.
Dolphin R, Wagley RA. 1977. Reading the annual report. *Financial Executive* **45**(6): 20–22.
Donelson DC, McInnis JM, Mergenthaler RD. 2012. Rules-based accounting standards and litigation. *The Accounting Review* **87**(4): 1247–1279.
Doran JS, Peterson DR, Price SM. 2010. Earnings conference call content and stock price: the case of REITS. *The Journal of Real Estate Finance and Economics* **45**(2): 402–434.
Dumais ST. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers* **23**(2): 229–236.
Durnev A, Mangen C. 2011. The real effects of disclosure tone: evidence from restatement. Available at SSRN: http://dx.doi.org/10.2139/ssrn.1650003
Engelberg J. 2008. Costly information processing: evidence from earnings announcements. In *AFA 2009 San Francisco Meetings*.
Engelberg JE, Reed AV, Ringgenberg MC. 2012. How are shorts informed? Short sellers, news, and information processing. *Journal of Financial Economics* **105**(2): 260–278.
Farrell BJ, Cobbin DM. 2000. A content analysis of codes of ethics from fifty-seven national accounting organizations. *Business Ethics: A European Review* **9**(3): 180–190.
Feldman R, Govindaraj S, Livnat J, Segal B. 2008. The incremental information content of tone change in management discussion and analysis. NYU Working Paper No. 2451/27580. Available at SSRN: http://ssrn.com/abstract=1280743
Feldman R, Govindaraj S, Livnat J, Segal B. 2010. Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* **15**(4): 915–953.
Ferguson D. 1997. Parsing financial statements efficiently and accurately using C and Prolog. Paper read at *Fifth International Conference on the Practical Application of Prolog*, London.
Ferguson NJ, Philip D, Lam HYT, Guo JM. 2014. Media content and stock returns: the predictive power of press. In *Midwest Finance Association 2013 Annual Meeting*.
Ferris SP, Hao GQ, Liao M. 2013. The effect of issuer conservatism on IPO pricing and performance. *Review of Finance* **7**(3): 993–1027.
Fisher IE. 2004. On the structure of financial accounting standards to support digital representation, storage, and retrieval. *Journal of Emerging Technologies in Accounting* **8**(3): 139–164.
Fisher IE. 2007. A prototype system for the temporal reconstruction of financial accounting standards. *International Journal of Accounting Information Systems* **8**(3): 139–164.
Fisher IE, Garnsey MR. 2006. The semantics of change as revealed through an examination of financial accounting standards amendments. *Journal of Emerging Technologies in Accounting* **3**(1): 41–59.
Fisher IE, Garnsey MR. 2010. Improving information retrieval from accounting documents: a prototype digital thesaurus for employee benefits. In *2010 AAA Midyear Meeting of the Information Systems and the Strategic and Emerging Technologies Sections*, Clearwater, FL.
Fisher IE, McEwen RA. 2009. On a logical structure for the authoritative accounting literature: a discussion of the FASB's codification structure. *Issues in Innovation* **3**(1): 32–56.
Fisher IE, Garnsey MR, Goel S, Tam K. 2010. The role of text analytics and information retrieval in the accounting domain. *Journal of Emerging Technologies in Accounting* **7**(1): 1–24.

DOI: 10.1002/isaf

Folsom D, Hribar P, Mergenthaler R, Peterson K. 2013. Principles-based standards and earnings attributes. Management Science, Forthcoming. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2046190
Francis J, Schipper K, Vincent L. 2002. Expanded disclosures and the increased usefulness of earnings announcements. *The Accounting Review* **77**(3): 515–546.
Frazier KB, Ingram RW, Tennyson BM. 1984. A methodology for the analysis of narrative accounting disclosures. *Journal of Accounting Research* **22**(1): 318–331.
Fung GPC, Yu JX, Lam W. 2002. News sensitive stock trend prediction. Paper read at *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD*), Taipei, Taiwan.
Fung GPC, Yu JX, Wai L. 2003. Stock prediction: integrating text mining approach using real-time news. Paper read at *IEEE International Conference on Computational Intelligence for Financial Engineering*.
Fung GPC, Yu JX, Lu H. 2005. The predicting power of textual information on financial markets. *IEEE Intelligent Informatics Bulletin* **5**(1): 1–10.
Gangolly JS. 1989. A computational view of financial accounting standards. In Artificial Intelligence in Accounting and Auditing, Vasarhelyi MA (ed.). Markus Wiener Publishing, Inc.: New York; 101–125.
Gangolly J. 2008. A preliminary examination of differences in language between U.S. GAAP and international GAAP. Paper read at *17th Annual Research Workshop on Artificial Intelligence and Emerging Technologies in Accounting Auditing and Tax*, Anaheim, CA.
Gangolly J, Tam K. 2000. On lexical acquisition for the financial reporting domain: preliminary results of the analysis of year 2000 EDGAR filings. Paper read at *Eleventh Annual Research Workshop on*: *Artificial Intelligence and Emerging Technologies in Accounting*, *Auditing and Tax*, San Antonio, TX.
Gangolly J, Wu Y. 2000. On the automatic classification of accounting concepts: preliminary results of the statistical analysis of term-document frequencies. *The New Review of Applied Expert Systems and Emerging Technologies* **6**: 81–88.
Gangolly JS, Hedley TP, Wong CT. 1991. Semantic knowledge bases for financial accounting standards. *Expert Systems with Applications* **3**(1): 117–128.
Garcia D. 2014. Sentiment during recessions. *The Journal of Finance* **68**(3): 1267–1300.
Garnsey MR. 2006a. Automatic classification of financial accounting concepts. *Journal of Emerging Technologies in Accounting* **3**(1): 21–39.
Garnsey MR. 2006b. Change in accounting language: a preliminary examination of accounting pronouncements. In *Fifteenth Annual Research Workshop on Artificial Intelligence and Emerging Technologies in Accounting*, *Auditing and Tax*, Washington, DC.
Garnsey MR. 2008. A preliminary examination of differences in language between U.S. GAAP and international GAAP. Paper read at *17th Annual Research Workshop on Artificial Intelligence and Emerging Technologies in Accounting Auditing and Tax*, Anaheim, CA.
Garnsey MR. 2009. A comparison of language in the codification and the FASB and its predecessor's pronouncements. In *AAA Information Systems Section Midyear Meeting*, Charleston, SC.
Garnsey MR, Fisher IE. 2008. Appearance of new terms in accounting language: a preliminary examination of accounting pronouncements and financial statements. *Journal of Emerging Technologies in Accounting* **5**(1): 17–36.
Garnsey MR, Hotaling AW. 2011. Improving identification of search terms in the FARS database. *Review of Business Information Systems* **11**(1): 45–56.
Garnsey MR, O'Neill J, Stokes L. 2009. FARS database searching: providing potential search terms to students. *Accounting Educators Journal* **30**: 69–90.
Gerde J, Jr. 2003. EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database. *Decision Support Systems* **35**(1): 7–29.
Gibson C, Schroeder N. 1990. Readability of management's discussion and analysis. *Accounting Horizons* **4**(4): 78–87.
Gidofalvi G. 2001. Using news articles to predict stock price movements. University of California, San Diego, Department of Computer Science and Engineering.
Gilbert E, Karahalios K. 2010. Widespread worry and the stock market. Paper read at *4th International AAAI Conference on Weblogs and Social Media*.
Goel S. 2008. Qualitative information in annual reports and the detection of corporate fraud: a natural language processing perspective. Doctoral dissertation, University at Albany .
Goel S, Gangoly J, Faerman S, Uzuner O. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* **7**(1): 25–46.
Goldberg L. 1965. An Inquiry into the Nature of Accounting. AAA: Sarasota, FL.

DOI: 10.1002/isaf

Govindarajan V. 1980. The objectives of financial statements: an empirical study of the use of cash-flow and earnings by security analysts. *Accounting, Organizations and Society* **5**(4): 383–392.
Grant GH, Conlon SJ. 2006. EDGAR extraction system: an automated approach to analyze employee stock option disclosures. *Journal of Information Systems* **20**(2): 119–142.
Groth SS, Muntermann J. 2009. Supporting investment management processes with machine learning techniques. Paper read at *The Internationale Tagung Wirtschaftsinformatik*, Vienna, Austria.
Groth SS, Muntermann J. 2011. An intraday market risk management approach based on textual analysis. *Decision Support Systems* **50**(4): 680–691.
Gu B, Konana P, Rajagopalan B, Chen H. 2007. Competition among virtual communities and user valuation: the case of investing-related communities. *Information Systems Research* **18**(1): 68–85.
Hagenau M, Liebmann M, Neumann D. 2013. Automated news reading: stock price prediction based on financial news using context-capturing features. *Decision Support Systems and Electronic Commerce* **55**(3): 685–697.
Hanley K, Hoberg G. 2008. Strategic disclosure and the pricing of initial public offerings. Working paper, University of Maryland.
Hanley K, Hoberg G. 2010. The information content of IPO prospectuses. *Review of Financial Studies* **23**(7): 2821–2864.
Harmon G. 1986. Information measurement: approaches, problems, prospects. Paper presented at the *Proceedings of the American Society for Information Science*.
Healy PM. 1977. Can you understand the footnotes to financial statements? *Accountants Journal* **56**(July): 219–222.
Henry E. 2006. Market reaction to verbal components of earnings press releases: event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting* **3**(1): 1–19.
Henry E. 2008. Are investors influenced by how earnings press releases are written? *Journal of Business Communication* **45**(4): 363–407.
Henry E, Leone AJ. 2010. Measuring qualitative information in capital markets research. Available at SSRN: http://dx.doi.org/10.2139/ssrn.1470807
Hoberg G, Phillips G. 2009. Product market synergies and competition in mergers and acquisitions. *Review of Financial Studies* **23**(10): 3773–3811.
Hofstedt TR. 1976. Behavioral accounting research: pathologies, paradigms and prescriptions. *Accounting, Organizations and Society* **1**(1): 43–58.
Holley CL, Early J. 1980. Are financial statements easy to read? *The Woman CPA* **42**(April): 9–13.
Hooghiemstra R. 2010. Letters to the shareholders: a content analysis comparison of letters written by CEOs in the United States and Japan. *The International Journal of Accounting* **45**(3): 275–300.
Hooks KL, Moon JE. 1993. A classification scheme to examine management discussion and analysis compliance. *Accounting Horizons* **7**(2): 41–59.
Hoskin R, Hughes J, Ricks W. 1986. Evidence on the incremental information content of additional firm disclosures made concurrently with earnings. *Journal of Accounting Research* **24**(3): 1–32.
Huang A, Zang A, Zheng R. 2010a. Informativeness of text in analyst reports: a naive Bayes machine learning approach. Working paper, Hong Kong University of Science & Technology.
Huang CJ, Liao JJ, Yang DX, Chang TY, Luo YC. 2010b. Realization of a news dissemination agent based on weighted association rules and text mining techniques. *Expert Systems with Applications* **37**(9): 6049–6413.
Huang KW, Li ZL. 2011. A multi-label text classification algorithm for labeling risk factors in SEC form 10-K. *ACM Transactions on Management Information Systems (TMIS)* **2**(3): 1–19.
Huang X. 2011. Text analysis of earnings press releases. Doctoral dissertation, University of California, Irvine, ProQuest Dissertations & Theses Database.
Huang X, Teoh SH, Zhang Y. 2014. Tone management. *The Accounting Review* **89**(3): 1083–1113.
Humpherys SL, Moffitt KC, Burns MB, Burgoon JK, Felix WF. 2011. Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems* **50**(3): 585–594.
Hussainey K, Schleicher T, Walker M. 2003. Undertaking large-scale disclosure studies when AIMR-FAF ratings are not available: the case of prices leading earnings. *Accounting and Business Research* **33**(4): 275–294.
Ijiri Y, Kinard JC, Putney FB. 1968. An integrated evaluation system for budget forecasting and operating performance with a classified budgeting bibliography. *Journal of Accounting Research* **6**: 1–28.
Ingram RW, Frazier KB. 1980. Environmental performance and corporate disclosure. *Journal of Accounting Research* **18**(2): 614–622.
Jacobs PS, Rau LF. 1990. SCISOR: Extracting information from on-line news. *Communications of the ACM* **33**(11): 88–97.

DOI: 10.1002/isaf

Jarvenpaa S, Ives B. 1990. Information technology and corporate strategy: a view from the top. *Information Systems Research* **1**(4): 351–376.
Jegadeesh N, Wu AD. 2013. Word power: a new approach for content analysis. *Journal of Financial Economics* **110**(3): 712–729.
Jiang W. 2005. Intelligent day trading agent: a natural language processing approach to financial information analysis. Doctoral dissertation, Rutgers, The State University of New Jersey, ProQuest Dissertations & Theses Database.
Jin F, Self N, Saraf P, Butler P, Wang W, Ramakrishnan N. 2013. Forex-foreteller: currency trend modeling using news articles. Paper read at *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL.
Jones MJ. 1988. A longitudinal study of the readability of the chairman's narratives in the corporate reports of a U.K. company. *Accounting and Business Research* **18**(72): 297–305.
Jones MJ, Shoemaker PA. 1994. Accounting narratives: a review of empirical studies of content and readability. *Journal of Accounting Literature* **13**(1): 142–184.
Kamaruddin SS, Hamdan AR, Bakar AA. 2007. Text mining for deviation detection in financial statements. Paper read at *International Conference on Electrical Engineering and Informatics*, 17–19 June, Bandung, Indonesia.
Kasznik R, Lev B. 1995. To warn or not to warn: management disclosures in the face of an earnings surprise. *The Accounting Review* **70**(1): 113–134.
Kearney C, Liu S. 2014. Textual sentiment in finance: a survey of methods and models. *International Review of Financial Analysis* **33**: 171–185.
Keila PS, Skillicorn DB. 2005. Detecting unusual and deceptive communication in email. In *CASCON '05 Proceedings of the 2005 Conference of the Centre for Advanced Studies on Collaborative Research*. IBM Press; 17–20.
Kelly N. 2014. *Why machines alone cannot solve the world's translation problem*. http://www.huffingtonpost.com/nataly-kelly/why-machines-alone-cannot-translation_b_4570018.html?utm_hp_ref=technology&ir=Technology (accessed 24/12/2015).
Kelly-Newton L. 1980. A sociological investigation of the U.S.A. mandate for replacement cost disclosures. *Accounting, Organizations and Society* **5**(3): 311–321.
Kimbrough MD, Wang IY. 2014. Are seemingly self-serving attributions in earnings press releases plausible? Empirical evidence. *The Accounting Review* **89**(2): 635–667.
Klamm BK, Watson MW. 2009. SOX 404 reported internal control weaknesses: a test of COSO framework components and information technology. *Journal of Information Systems* **23**(2): 1–23.
Kothari SP, Li X, Short JE. 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: a study using content analysis. *The Accounting Review* **84**(5): 1639–1670.
Kravet T, Muslu V. 2011. Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies* **18**(4): 1088–1122.
Kumar RB, Kumar BS, Prasad CSS. 2012. Financial news classification using SVM. *International Journal of Scientific and Research Publications* **2**(3): 1–6.
Larcker DF, Zakolyukina AA. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* **50**(2): 495–540.
Lavrenko V, Schmill M, Lawrie D, Ogilvie P, Jensen D, Allan J. 2000. Language models for financial news recommendation. Paper read at *9th International Conference on Information and Knowledge Management*.
Lawrence AN. 2011. Individual investors and financial disclosure. *Journal of Accounting and Economics* **56**(1): 130–147.
Lebar MA. 1982. A general semantics analysis of selected sections of the 10-K, the annual report to shareholders, and the financial press release. *The Accounting Review* **57**(1): 176–189.
Lee YJ. 2010. The effect of quarterly report readability on information efficiency of stock prices. *Contemporary Accounting Research* **29**(4): 1137–1170.
Lehavy R, Li F, Merkley K. 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review* **86**(3): 1087–1115.
Leinemann C, Schlottmann F, Seese D, Stuempert T. 2001. Automatic extraction and analysis of financial data from the EDGAR database. *South African Journal of Information Management* **3**(2). doi:10.4102/sajim.v3i2.127.
Levine C, Smith M. 2006. Critical accounting policy disclosures. *Journal of Accounting, Auditing & Finance* **26**(1): 39–76.
Li F. 2006. Do stock market investors understand the risk sentiment of corporate annual reports? Doctoral dissertation, University of Michigan, Ann Arbor, MI.

DOI: 10.1002/isaf

Li F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* **45**(2–3): 221–247.
Li F. 2010a. The determinants and information content of the forward-looking statements in corporate filings—a naive Bayesian machine learning approach. Paper read at *AAA 2009 Financial Accounting and Reporting Section*.
Li F. 2010b. The information content of forward-looking statements in corporate filings—a naive Bayesian machine learning approach. *Journal of Accounting Research* **48**(5): 1049–1102.
Li F. 2010c. Textual analysis of corporate disclosures: a survey of the literature. *Journal of Accounting Literature* **29**: 143–165.
Li F, Lundholm R, Minnis M. 2011. The impact of perceived competition on the profitability of investments and future stock returns. Working paper, University of Michigan and University of British Columbia.
Li Q, Wang TJ, Li P, Liu L, Gong Q, Chen Y. 2014a. The effect of news and public mood on stock movements. *Information Sciences* **278**: 826–840.
Li X, Huang X, Deng X, Zhu S. 2014b. Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. *Neurocomputing* **142**: 228–238.
Liddy ED. (2001) Natural language processing. In Encyclopedia of Library and Information Science, 2nd edn, Drake MA (ed.). Marcel Decker: New York.
Linsley PM, Shrives PJ. 2006. Risk reporting: a study of risk disclosures in the annual reports of UK companies. *The British Accounting Review* **38**(4): 387–404.
Loughran T, McDonald B. 2009. Plain English, readability, and 10-K filings. Working paper, University of Notre Dame.
Loughran T, McDonald B. 2010. Measuring readability in financial text. Working paper, The University of Notre Dame.
Loughran T, McDonald B. 2011a. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* **66**(1): 35–65.
Loughran T, McDonald B. 2011b. Barron's red flags: do they actually work? *Journal of Behavioral Finance* **12**(2): 90–97.
Loughran T, McDonald B. 2013. IPO first-day returns, offer price revisions, volatility, and Form S-1 language. *Journal of Financial Economics* **109**(2): 307–326.
Loughran T, McDonald B. 2014. Measuring readability in financial disclosures measuring readability in financial disclosures. *The Journal of Finance* **69**(4): 1643–1671.
Loughran T, McDonald B, Yun H. 2009. A wolf in sheep's clothing: the use of ethics-related terms in 10-K reports. *Journal of Business Ethics* **89**(1): 39–49.
Lu YC, Shen CH, Wei YC. 2013. Revisiting early warning signals of corporate credit default using linguistic analysis. *Pacific-Basin Finance Journal* **24**: 1–21.
Lugmayr A, Gossen G. 2012. Evaluation of methods and techniques for language based sentiment analysis for DAX 30 stock exchange—a first concept of a 'LUGO' sentiment indicator. Paper read at *SAME 2012—5th International Workshop on Semantic Ambient Media Experience*.
Luo X, Zhang J, Duan W. 2013. Social media and firm equity value. *Information Systems Research* **24**(1): 146–163.
Lytinen SL, Gershman A. 1986. ATRANS: automatic processing of money transfer messages. In *Proceedings of the 5th National Conference on Artificial Intelligence*, *Volume 2*: *Engineering*, Philadelphia, PA, 11–15 August.
Magnusson C, Arppe A, Eklund T, Back B, Vanharanta H, Visa A. 2005. The language of quarterly reports as an indicator of change in the company's financial status. *Information & Management* **42**(4): 561–574.
Mahajan A, Dey L, Haque M. 2008. Mining financial news for major events and their impacts on the market. Paper read *at IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent AgentTechnology*, Sydney, Australia.
Malo P, Sinha A, Takala P, Korhonen P, Wallenius J. 2014. Good debt or bad debt: detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* **65**(4): 782–796.
Masli A, Richardson VJ, Watson MW, Zmud RW. 2009. CEO, CFO & CIO engagement in information technology management: the disciplinary effects of Sarbanes-Oxley information technology material weaknesses. Paper read at *University of Waterloo Centre for Information Integrity & Information Systems Assurance 6th Bi-Annual Research Symposium*.
Matsumoto D, Pronk M, Roelofsen E. 2011. What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review* **86**(4): 1383–1414.
McConnell D, Haslem JA, Gibson VR. 1986. The president's letter to stockholders: a new look. *Financial Analysts Journal* **42**(5): 66–70.

DOI: 10.1002/isaf

Mendonca J. 2009. Analyst statements, stockholder reactions, and banking relationships: do analysts' words matter? Doctoral dissertation, The University of Texas at Austin, ProQuest Dissertations & Theses Database.
Merkl-Davies DM, Brennan NM. 2007. Discretionary disclosure strategies in corporate narratives: incremental information or impression management? *Journal of Accounting Literature* **26**: 116–196.
Merkley K. 2011. More than numbers: R&D-related disclosure and firm performance. Dissertation, The University of Michigan.
Merkley KJ. 2014. Narrative disclosure and earnings performance: evidence from R&D disclosures. *The Accounting Review* **89**(2): 725–757.
Meyer M, Rigsby JT. 2001. A descriptive analysis of the content and contributors of behavioral research in accounting 1989–1998. *Behavioral Research in Accounting* **13**(1): 253–278.
Miller B. 2010. The effects of reporting complexity on small and large investor trading. *The Accounting Review* **85**(6): 2107–2143.
Miller GS. 2002. Earnings performance and discretionary disclosure. *Journal of Accounting Research* **40**(1): 173–203.
Mirakur Y. 2011. Risk disclosure in SEC corporate filings. *Wharton Research Scholars Journal*: paper 85.
Mittermayer MA. 2004. Forecasting intraday stock price trends with text mining techniques. Paper read at *37th Annual Hawaii International Conference on System Sciences*, Hawaii.
Mittermayer MA, Knolmayer GF. 2006. NewsCATS: a news categorization and trading system. Paper read at *6th International Conference on Data Mining* (*ICDM*).
Moffitt K, Burns MB. 2009. What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. Paper read at *Fifteenth Americas Conference on Information Systems* (*AMCIS*), San Francisco, CA.
Moffitt K, Burns M. 2011. Using lexical bundles to discriminate between fraudulent and non-fraudulent financial reports. Working paper.
Mohan S. 2006. Disclosure quality and its effect on litigation risk. Working paper. http://ssrn.com/abstract=956499 or http://dx.doi.org/10.2139/ssrn.956499 (accessed 27 December 2015).
Müller A, Dörre J, Gerstl P, Seiffert R. 1999. The TaxGen framework: automating the generation of a taxonomy for a large document collection. Paper read at *32nd Annual Hawaii International Conference on System Sciences* (*HICSS*), Hawaii.
Muslu V, Rahhakrishnan S, Subramanyam KR, Lim D. 2010. Forward-looking disclosures in the MD&A and the financial information environment. *Management Science* **61**(5): 931–948.
Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL. 2014. Text mining for market prediction: a systematic review. *Expert Systems with Applications* **41**(16): 7653–7670.
Nelson KK, Pritchard AC. 2007. Litigation risk and voluntary disclosure: the use of meaningful cautionary language. Paper read at *2nd Annual Conference on Empirical Legal Studies Paper*.
O'Leary DE. 2011. Blog mining-review and extensions: 'from each according to his opinion'. *Decision Support Systems* **51**(4): 821–830.
O'Leary, DE. 2012. Knowledge discovery for continuous financial assurance using multiple types of digital information. In *World Conference on Continuous Auditing and Repo*, Rutgers University.
O'Leary DE. 2013. Big data', the 'Internet of Things' and the 'Internet of Signs'. *Intelligent Systems in Accounting, Finance and Management* **20**(1): 53–65.
Othman IW, Hasan H, Tapsir R, Rahman NA, Tarmuji I, Majdi S, Masuri SA, Omar N. 2012. Text readability and fraud detection. Paper read at *IEEE Symposium on Business*, *Engineering and Industrial Applications* (*ISBEIA 2012*).
Ozik G, Sadka R. 2010. Media and investment management. Available at SSRN: http://dx.doi.org/10.2139/ssrn.1633705
Pashalian S, Crissy WJE. 1952. Corporate annual reports are difficult, dull reading, human interest value low. *Journal of Accountancy* **94**(2): 215–219.
Pava ML, Epstein MJ. 1993. How good is MD&A as an investment tool? *Journal of Accountancy* **175**(3): 51–53.
PCAOB. 2010. Auditing Standard No. 12: Identifying and Assessing Risks of Material Misstatement. http://pcaobus.org/Standards/Auditing/Pages/Auditing_Standard_12.aspx (accessed 27 December 2015).
Peramunetilleke D, Wong RK. 2002. Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications* **24**: 131–139.
Previts G, Brown R. 1993. The development of government accounting: a content analysis of the *Journal of Accountancy*, 1905–1989. *The Accounting Historians Journal* **20**(2): 119–138.
Previts G, Bricker R, Robinson T, Young S. 1994. A content analysis of sell-side analyst company reports. *Accounting Horizons* **8**(2): 55–70.

DOI: 10.1002/isaf

Price SM, Doran JS, Peterson DR, Bliss BA. 2012. Earnings conference calls and stock returns: the incremental informativeness of textual tone. *Journal of Banking & Finance* **36**(4): 992–1011.
Purda L, Skillicorn D. 2014. Accounting variables, deception, and a bag of words: assessing the tools of fraud detection. *Contemporary Accounting Research* **32**(3): 1193–1223.
Qiu XY, Jiang S, Deng K. 2013. Automatic assessment of information disclosure quality in Chinese annual reports. Natural Language Processing and Chinese Computing, Zhou G, Li J, Zhao D, Feng Y (eds). Springer: Berlin; 288–298.
Qiu XY, Srinivasan P, Hu Y. 2014. Supervised learning models to predict firm performance with annual reports: an empirical study. *Journal of the Association for Information Science and Technology* **65**(2): 400–413.
Rachlin G, Last M, Alberg D, Kandel A. 2007. ADMIRAL: a data mining based financial trading system. Paper read at *IEEE Symposium on Computational Intelligence and Data Mining* (*CIDM*).
Rees L, Twedt B. 2012. Reading between the lines: an empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy* **31**(1): 1–21.
Rezaee Z, Olibe KO, Minmier G. 2003. Improving corporate governance: the role of audit committee disclosures. *Managerial Auditing Journal* **18**(6–7): 530–537.
Rogers JL, Van Buskirk A, Zechman SLC. 2011. Disclosure tone and shareholder litigation. *The Accounting Review* **86**(6): 2155–2183.
Rogers RK, Grant J. 1997. Content analysis of information cited in reports of sell-side financial analysts. *Journal of Financial Statement Analysis* **3**: 17–31.
Rutherford BA. 2005. Genre analysis of corporate annual report narratives a corpus linguistics-based approach. *Journal of Business Communication* **42**(4): 349–378.
Sadique S, In F, Veeraraghavan M. 2008. The impact of spin and tone on stock returns and volatility: evidence from firm-issued earnings announcements and the related press coverage. Working paper, Monash University.
Sadique S, In F, Veeraraghavan M, Wachtel P. 2013. Soft information and economic activity: evidence from the Beige Book. *Journal of Macroeconomics* **37**: 81–92.
Schroeder N, Gibson C. 1990. Readability of management's discussion and analysis. *Accounting Horizons* **4**(4): 78–87.
Schumaker RP, Chen H. 2009. Textual analysis of stock market prediction using breaking financial news: the AZFinText system. *ACM Transactions on Information Systems* **27**(2): 12.
Schumaker RP, Zhang Y, Huang C, Chen H. 2012. Evaluating sentiment in financial news articles. *Decision Support Systems* **53**(3): 458–464.
Seo YW, Giampapa J, Sycara K. 2002. Text classification for intelligent portfolio management (no. CMU-RI-TR-02–14). Carnegie Mellon University, Robotics Institute: Pittsburgh, PA.
Shon J. 2003. The relation between earnings surprises and discretionary disclosure behavior in periods with short-term price declines. Working paper, University of Chicago.
Singhvi SS. 1968. Corporate disclosure through annual reports in the United States of America and India. *The Journal of Finance* **23**(3): 551–552.
Sinha NR. 2010. Underreaction to news in the US stock market. Working paper, University of Illinois at Chicago.
Sklar H. 2011. *Risk Asssessments: The Most Important Effort You're Doing All Wrong*. http://www.forbes.com/sites/howardsklar/2011/11/22/risk-assessments-the-most-important-effort-youre-doing-all-wrong/ (accessed 27 December 2015).
Smales LA. 2014. News sentiment and the investor fear gauge. *Finance Research Letters* **11**(2): 122–130.
Smith JE, Smith NP. 1971. Readability: a measure of performance of the communication function of financial reporting. *The Accounting Review* **46**(3): 552–556.
Smith M, Taffler R. 1992a. Readability and understandability: different measures of the textual complexity of accounting narrative. *Accounting, Auditing & Accountability Journal* **5**(4): 84–98.
Smith M, Taffler RJ. 1992b. The chairman's statement and corporate financial performance. *Accounting and Finance* **32**(2): 75–90.
Smith M, Taffler RJ. 2000. The chairman's statement: a content analysis of discretionary narrative disclosures. *Accounting, Auditing & Accountability Journal* **13**(5): 624–626.
Solomon DH, Soltes E, Sosyura D. 2014. Winners in the spotlight: media coverage of fund holdings as a driver of flows. *Journal of Financial Economics* **113**(1): 53–72.
Soni A, van Eck NJ, Kaymak U. 2007. Prediction of stock price movements based on concept map information. Paper read at *IEEE Symposium on Computational Intelligence in Multicriteria Decision Making*, Honolulu, HI.
Soper FJ, Dolphin R. 1964. Readability and corporate annual reports. *The Accounting Review* **39**(2): 358–362.

DOI: 10.1002/isaf

Steele A. 1982. The accuracy of chairman's non-quantified forecasts: an exploratory study. *Accounting and Business Research* **12**(47): 215–230.
Still MD. 1972. The readability of chairman's statements. *Accounting and Business Research* **3**(9): 36–39.
Summers SL, Sweeney JT. 1998. Fraudulently misstated financial statements and insider trading: an empirical analysis. *The Accounting Review* **73**(1): 131–146.
Swales GS, Jr. 1988. Another look at the president's letter to stockholders. *Financial Analysts Journal* **44**(2): 71–73.
Sydserff R, Weetman P. 1999. Methodological themes: a texture index for evaluating accounting narratives—an alternative to readability formulas. *Accounting, Auditing & Accountability Journal* **12**(4): 459–488.
Sydserff R, Weetman P. 2002. Developments in content analysis: a transitivity index and scores. *Accounting, Auditing & Accountability Journal* **15**(4): 523–545.
Tennyson BM, Ingram RW, Dugan MT. 1990. Assessing the information content of narrative disclosures in explaining bankruptcy. *Journal of Business Finance & Accounting* **17**(3): 390–410.
Tetlock PC. 2007. Giving content to investor sentiment: the role of media in the stock market. *The Journal of Finance* **62**(3): 1139–1168.
Tetlock PC. 2011. All the news that's fit to reprint: do investors react to stale information? *Review of Financial Studies* **24**(5): 1481–1512.
Tetlock C, Saar-Tsechansky M, Macskassy S. 2008. More than words: quantifying language to measure firms' fundamentals. *The Journal of Finance* **63**: 1437–1467.
Thomas JD, Sycara K. 2002. Integrating genetic algorithms and text learning for financial prediction. In *Proceedings of the GECCO-2000 Workshop on Data Mining with Evolutionary Algorithms*; 72–75.
Tung WL, Quek C, Cheng P. 2004. GenSo-EWS: A novel neural-fuzzy based early warning system for predicting bank failures. *Neural Networks* **17**(4): 567–587.
Twedt B, Rees L. 2012. Reading between the lines: an empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy* **31**(1): 1–21.
Vu TT, Chang S, Ha QT, Collier N. 2012. An experiment in integrating sentiment features for tech stock prediction in Twitter. Paper read at *Workshop on Information Extraction and Entity Analytics on Social Media Data*, Mumbai, India.
Wächter T, Fabian G, Schroeder M. 2011. DOG4DAG: semi-automated ontology generation in OBO-Edit and Protégé. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*. ACM; New York; 119–120.
Wang B, Guo X. 2012. Online recruitment information as an indicator to appraise enterprise performance. *Online Information Review* **36**(6): 903–918.
Wang B, Wang X. 2012. Deceptive financial reporting detection: a hierarchical clustering approach based on linguistic features. *Procedia Engineering* **29**: 3392–3396.
Wang B, Huang H, Wang X. 2012a. A novel text mining approach to financial time series forecasting. *Neurocomputing* **83**: 136–145.
Wang H, Li L, Cao J. 2012b. Lexical features in corporate annual reports: a corpus-based study. *European Journal of Business and Social Sciences* **1**(9): 55–71.
Wuthrich B, Cho V, Leung S, Permunetilleke D, Sankaran K, Zhang J. 1998. Daily stock market forecast from textual web data. Paper read at *IEEE International Conference on Systems*, *Man*, *and Cybernetics*, San Diego, CA.
Yen A, Hirst E, Hopkins P. 2007. A content analysis of the comprehensive income exposure draft comment letters. *Research in Accounting Regulation* **19**: 53–79.
You H, Zhang X. 2009. Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies* **14**(4): 559–586.
Yu Y, Duan W, Cao Q. 2013. The impact of social and conventional media on firm equity value: a sentiment analysis approach. *Decision Support Systems* **55**(4): 919–926.
Zhai Y, Hsu A, Halgamuge SK. 2007. Combining news and technical indicators in daily stock price trends prediction. In Advances in Neural Networks – ISNN 2007: 4th International Symposium on Neural Networks, ISSN 2007, Nanjing, China, June 3–7, 2007, Proceedings, Part III, Liu D, Fei S, Zhang H, Sun C (eds). Lecture Notes in Computer Science, vol. **4493**. Springer: Berlin; 1087–1096.

DOI: 10.1002/isaf