ANDY MONIZ

Textual Analysis of Intangible Information

Textual Analysis of Intangible Information

Textual Analysis of Intangible Information

Tekstuele analyse van immateriële informatie

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Thursday, 16 June 2016, at 15:30 hrs

by

ANDREW JULIAN MONIZ
born in London, the United Kingdom.

Erasmus University Rotterdam

Doctoral Committee:

Promotors:	Prof.dr. C.B.M. van Riel Prof.dr.ir. F.M.G. de Jong
Other members:	Prof.dr. J.P. Cornelissen Prof.dr. P.P.M.A.R. Heugens Prof.dr. C.G. Koedijk
Copromotor:	Dr. G.A.Y.M. Berens

Erasmus Research Institute of Management – ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://repub.eur.nl/pub

ERIM PhD Series in Research in Management
ERIM reference number: EPS-2016-393-ORG
ISBN 978-90-5892-451-3

Design: Courtesy of Hugo and Florence Moniz

This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

Acknowledgements

Writing a Ph.D. thesis in interdisciplinary fields is challenging and demanding yet an exciting and rewarding experience. First and foremost, I would like to express my deepest gratitude to my supervisors for their continual words of encouragement and motivation which enabled me to pursue this research. Cees, on the very first occasion we met you ended the conversation by saying ‘when do we start?’ Ever since then, you have been a source of inspiration. Without your interest and support, I would not have been able to complete this thesis within a relatively short period of time. Guido, thank you so much for your patience, help and dedication. Your insightful comments and constructive criticisms at different stages of my research were thought-provoking and helped challenge and refine my ideas. Working on this research whilst holding down a full-time job posed many challenges. Thank you for accommodating my unusual working practices with long-periods of ‘radio silence’. Franciska, I have benefitted tremendously from your insights and your rigorous approach to research. The skills I have learned from you have changed the way I address issues both in an academic and work context. Thank you for encouraging me to present at computer science conferences. Your feedback on papers, which I typically requested at unearthly hours to take advantage of Pacific Standard Time submission deadlines, was far beyond the call of duty.

I am grateful to my colleagues at UBS O’Connor (Braden Janowski and Bryan Smith) and APG Asset Management (Gerben de Zwart, Ronald van Dijk, Terhi Halme and Bas de Bree). If it wasn't for their support, encouragement and advice I may never have decided to pursue this area of research. Last but not least, none of this would have been possible without the love and patience of my family. Winnie, for supporting me throughout this endeavor and for taking care of the kids so often. Hugo and Florence, I dedicate this thesis to you to encourage you to follow your dreams.

London, June 2016

Andy Moniz

Contents

Chapter 1

"Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom" – Clifford Stoll

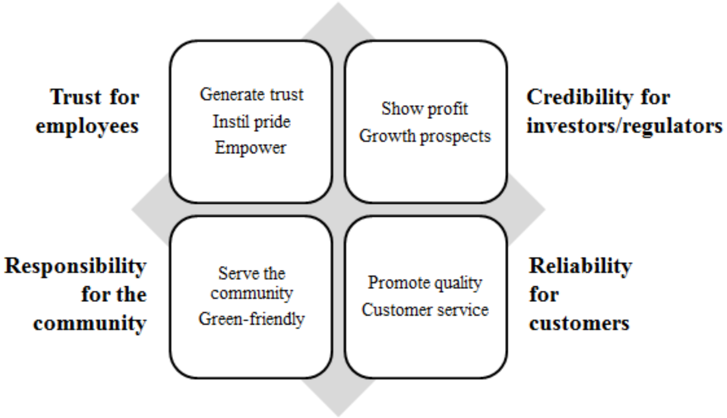
1.1 Introduction

In an efficient market, the value of a firm is equal to the value of its net assets minus its liabilities. This accounting identity is relatively straightforward to evaluate when physical assets (such as plants, property and equipment) represent a large proportion of a firm's balance sheet, yet can be more challenging to evaluate when a firm derives substantial value from intangible assets (Gu and Lev 2004; Lev et al. 2009; Barney 1991). Examples include corporate reputation (Fombrun and van Riel 1997), brand value (Madden et al. 2006), innovative efficiency (Hishleifer et al. 2013; Chan et al. 2001), human capital (Edmans 2011; Barney and Wright 1998) and organizational capital (Lev et al. 2009). In particular, the lack of physical substance associated with intangible assets, their opaque ownership rights and non-existent market prices, makes their measurement almost impossible (Gu and Lev 2004; Lev et al. 2009). Consequently, the conservative nature of international accounting standards limits firms from valuing and recording most types of intangible assets in their financial statements. Prior accounting literature even suggests that the lack of accurate accounting may have led to the 'systematic undervaluation of intangibles' and a sub-optimal allocation of resources (Lev 2001). Thus, the task of quantifying firms' intangible assets is important for corporate managers, policy makers, and investors seeking to resolve this 'value paradox' (see Skinner 2008a; Blaug and Lekhi 2009). Until accounting standards change, investors seeking to integrate intangible asset valuations into their decision making processes must seek alternative sources of information beyond a firm's own financial statements (Angelopoulos et al. 2012).

One alternative source of information is publicly available data published on the Internet. The Web has empowered users to create and share a wealth of information in the form of opinions, ideas and experiences (Gaines-Ross 2010). In contrast to accounting data which reside in a traditional row-column database, Web data are considered "unstructured". This is because the variety of text and multimedia content doesn't fit neatly into a standard database. In this thesis we employ Information Retrieval and Natural Language Processing (NLP) techniques to infer the measurement of a firm's intangible assets. We retrieve unstructured data in the form of authors' opinions from the Web and media news for a broad number of publicly listed companies. To decide whose opinions are relevant, we draw upon the concept of stakeholder theory (Freeman 1984; Jensen 2001). Stakeholders are commonly defined in organizational literature as *"any group or individual who can affect or is affected by the achievement of the organization's objectives"* (Freeman 1984). Primary stakeholders are those groups without whose continuing participation the corporation cannot survive as a going concern, which include a firm's employees, customers, suppliers and shareholders (Clarkson 1995; Orlitzky and Benjamin 2001), while secondary stakeholders refer to those groups who influence or affect the firm but who are not engaged in transactions with the firm and are not essential for its survival (such as journalists and non-governmental organizations). As illustrated in Figure 1.1, one implication of stakeholder theory is the view that the effective management of stakeholder relationships can mitigate the likelihood of negative regulatory and legislative action (Freeman 1984; Berman 1999), attract socially conscious consumers (Hillman and Kleim 2001) and increase firm performance by protecting and enhancing corporate reputation (Fombrun and Shanley 1990; Fombrun 2005; Freeman et al. 2007).

Figure 1.1: Illustrative example of the stakeholder perspective

The diagram below is adapted from Fombrun 1998 and depicts different stakeholder perspectives (employees, local communities, customers, investors and regulators).



We retrieve a variety of unstructured datasets to infer different stakeholder perspectives including corporate environmental sustainability disclosures (society's perspective), financial media news (an investor's perspective), central bank communications (a regulatory perspective), and social media (an employee's perspective).

1.2 Statistical textual analysis

The integration of intangible information into investment analysis requires tools to retrieve and categorize large volumes of text and dimension reduction techniques to analyse information across companies in a consistent way. Prior financial text mining studies have primarily relied upon a Naïve Bayesian model to classify documents (see Engelberg 2008). The Naïve Bayesian model is one of the simplest and most commonly used machine-learning algorithms for text classification (see Manning and Schütze 1999) and has been shown to be highly successful at classifying documents within the financial domain (Antweiler and Frank 2006). The model assumes that a document is generated by first choosing a topic z and then generating N number of words (w) independently from the conditional multinomial distribution $p(w|z)$. A document can be a sentence, a single paragraph, a webpage, an e-mail message, a newspaper article, or any written text. The model assumes that each document contains only one topic and the set of possible topics must be provided. The probability model is defined for a document d as:

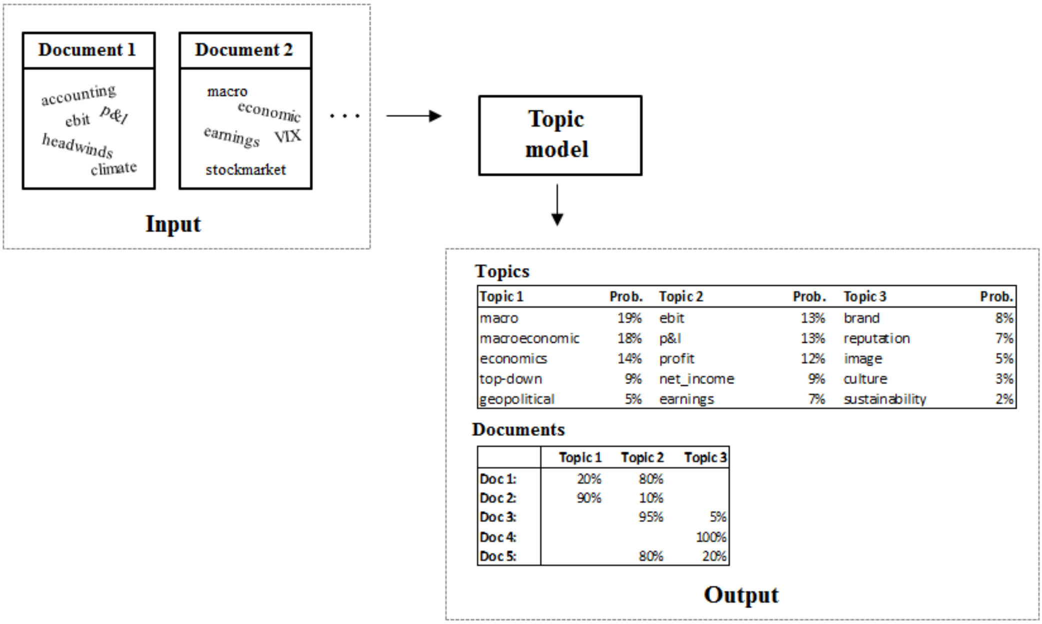
$$p(d) = \sum_{z} p(z) \prod_{n=1}^{N_d} p(w_n\,|z). \tag{1.1}$$

Despite the model's simplicity, it has several limitations for the robust modeling of documents (see Gimpel 2006). In particular, the model makes the simplifying assumption that the author of a document discusses only one topic in the text. While this assumption may be appropriate for the classification of accounting information such as 'earnings' news (see Tetlock 2008; Loughran and McDonald 2011), we start from the premise that the one topic assumption is less likely to be valid for the classification of intangible information due to the greater variety of topics discussed in text (Gurun et al. 2012).

To infer measures of intangible information we employ a probabilistic topic model. Topic models are statistical models that posit low-dimensional representations of data and provide an alternative approach to classify documents. The most common topic model is Latent Dirichlet Allocation (LDA) (Blei et al. 2003). The basic premise of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The model is based on the hypothesis that an author writing a document has certain topics in mind. To write about a topic requires picking a word with a certain probability from a pool of words from that topic. A document can then be represented as a mixture of different topics. These topics reflect the author's view of a document and his/her particular vocabulary. The input to a topic model is a set of documents with each document typically represented as a "bag-of-words". This simplifying assumption implies that the probability of each word in a document is independent of the word's context and position in the document (i.e. word order does not matter). The output of the LDA model is a set of topics together with probabilistic topic assignments for each document. Figure 1.2 provides an overview of a topic model pipeline. Document 1 contains a large proportion of terms from Topic 2 (which appear to be accounting related) while Document 2 contains a large proportion of terms from Topic 1 (which appear to be related to the economy). Thus, the goal of topic modeling is to automatically discover the latent topics from the collection of documents. The documents themselves are observed, while the topic structure - the topics, per-document topic distributions, and the per-document per-word topic assignments are hidden (Blei et al. 2003).

Figure 1.2: Illustrative example of the LDA topic model pipeline

The diagram below depicts the LDA topic model pipeline. The input of a topic model is a set of documents. Each document is represented as a bag-of-words. The output is a set of topic clusters together with topic assignments for each word. Topic labels are inferred from the words with the highest probability in each cluster.



Topic 1	Prob.	Topic 2	Prob.	Topic 3	Prob.
macro	19%	ebit	13%	brand	8%
macroeconomic	18%	p&l	13%	reputation	7%
economics	14%	profit	12%	image	5%
top-down	9%	net_income	9%	culture	3%
geopolitical	5%	earnings	7%	sustainability	2%

	Topic 1	Topic 2	Topic 3
Doc 1:	20%	80%	
Doc 2:	90%	10%	
Doc 3:		95%	5%
Doc 4:			100%
Doc 5:		80%	20%

In the standard implementation of LDA the topic clusters are unlabelled and must be manually annotated by the user. This is typically achieved by ranking the top terms for each topic cluster by their marginal probabilities *p(w|z)* (see Blei et al. 2003; Griffiths and Steyvers 2004). For instance, Topic 3 allocates the highest probabilities to the terms 'brand', 'reputation' and 'image', suggesting that this topic relates to discussions of a firm's intangible assets. Despite the potential simplicity of this approach, manual annotation can entail significant cognitive burden in interpretation, is prone to subjectivity, and lacks reproducibility (Lau et al. 2011; Newman et al. 2010). To address these concerns, this thesis examines a variant of LDA which integrates financial domain knowledge to steer the LDA model towards topics of interest (see Jagarlamudi et al. 2012 for details of the methodology). Specifically, we seed the LDA model with sets of words either from financial lexicons (see Loughran and McDonald 2011) or online databases (e.g. DBPedia and SPARQL endpoints) to improve both topic word distributions (by biasing topics to produce appropriate seed words) and to improve document-topic distributions (by biasing documents to select topics).

1.3 Aspect-level sentiment analysis

Sentiment analysis studies the phenomena of opinion, sentiment, evaluation, appraisal, attitude and emotion. An opinion can be defined as a *"judgment or belief not founded on certainty or proof*" (see Schouten and Frasincar 2015 for a review of the literature). In this sense, statements expressing an opinion are subjective, while factual statements are objective. Traditionally, aspect-level sentiment analysis has been researched for the classification of online user reviews of products and movies (Pang et al. 2002). Readers are often not only interested in the general sentiment towards an aspect but also a detailed opinion analysis for each of these aspects (Titov and McDonald 2008). Aspects are topics on which opinions are expressed. Aspects are important because without knowing them the opinions expressed are of limited use (Bagheri et al. 2013b).

The earliest attempts at aspect detection were based on an information extraction of frequently occurring nouns in text (see Hu and Liu 2004). Such approaches work well in detecting aspects that are strongly associated with a single noun (such as 'price' or 'quality'), but are less useful when aspects encompass many low frequency terms or are abstract (Brody and Elhadad 2010). In this regard, mining and interpreting opinions about companies is a harder and less understood problem than opinion mining for products and services. Stakeholder groups (ranging from consumers, investors, regulators, and local communities) may weigh aspects such as a company's product/service quality, management quality, and environmental sustainability differently. While the seminal work by van Riel (1995) provided a structural approach to measure reputation along different aspects, with the ever increasing amount of opinion published on the Web, there is a strong demand for automatically retrieving and summarizing the opinions expressed in natural language text. In the extreme case of social media, with 500 million tweets published per day and more than 750 million daily active Facebook users, manual approaches to reputation analysis lose their feasibility (Peetz 2015). Probabilistic topic models provide an unsupervised way to discover aspects in text.

Sentiment is orthogonal to opinion and is used to convey an evaluation of a topic in text (Kim and Hovy 2004). Consequently, text can fall into four quadrants - it can be subjective or objective, as well as with or without sentiment (Schouten and Frasincar 2015). Finding sentiment can be formally defined as finding the quadruple (*s; g; h; t*) (Liu 2012), where *s* represents the sentiment, *g* represents the target object for which the sentiment is expressed, *h* represents the holder (i.e., the one expressing the sentiment), and *t* represents the time at which the sentiment was expressed. The target can be an entity, such as the overall topic, or an aspect of an entity, which can be any characteristic or property of that entity (Schouten and Frasincar 2015). Depending on the specific domain, different tasks in sentiment analysis become important. For example, more formal language is often observed in newswire text compared to microblogging posts where informal language, slang, errors in spelling and grammar are more frequently observed (see Derczynski et al. 2013). The task of detecting the expression of emotion in natural language text can be considered as a refinement of sentiment analysis. The goal is to classify a piece of text according to a predefined set of basic emotions. Most sentiment analysis approaches employed by the extant financial literature are based upon counting the number of positive versus negative terms in the text using a pre-defined lexicon (known as a ‘term counting’ approach). These lexicons contain words tagged with a polarity (also called affective valence or semantic orientation) to indicate whether a word conveys positive or negative content. The two generally accepted lexicons employed by the financial studies are the Harvard IV-4 psychosocial dictionary (Stone 1966) and the LM neg word list (Loughran and McDonald 2011).

1.4 Goal and perspective of this thesis

This thesis investigates the problem of designing automated measures to infer intangible information for publicly listed firms from unstructured data and evaluates the benefits of their integration into investment processes. Whether text is incrementally useful to investors beyond ‘hard’ accounting information is still an open empirical question. Text might not provide independent information if investors use it to justify contemporaneous quantitative measures (Francis and Soffer 1997). Furthermore, investors may find text difficult to integrate into decision making processes because it may not be verifiable ex post compared to accounting data, comparable across companies, or easily converted into numerical inputs (Huang et al. 2014). Thus, the problem statement that motivates this thesis is:

How can an automated system infer qualitative aspects from unstructured data sets and aggregate it into actionable, valuable financial knowledge?

This thesis answers three research questions related to textual analysis in finance:

Question 1: How can we use online texts to infer intangible information for firms?

Question 2: How can we integrate intangible information into investment analysis?

Question 3: Is intangible information incremental to the prediction of firms’ earnings?

The analysis of unstructured data relates to three main tasks: i) the extraction and representation of textual information to proxy the intangible assets of a firm, ii) the aggregation of measures of intangible information with hard accounting information into actionable knowledge (investment analysis), and iii) the evaluation of the combined measures for the prediction of firms’ earnings. These three tasks are addressed in this thesis through an interdisciplinary approach. For the extraction and representation of intangible information we draw upon methodologies employed in Information Retrieval and Natural Language Processing literature. The representation of intangible information is realized by drawing upon constructs from the field of organizational studies. The aggregation and evaluation of information relies upon regression techniques developed in financial asset pricing literature.

1.5 Academic contribution of this thesis

The added-value of this thesis can be divided into contributions of new models to infer measures of intangible information and empirical analyses for predicting firms' earnings.

1.5.1 Models to infer intangible information

Despite a wealth of NLP literature describing methodologies to infer individuals' opinions and experiences (see Pang and Lee 2004), NLP applications to the financial domain appear to be relatively underresearched. While prior financial literature has investigated the merits of textual analysis in the context of companies' earnings conference calls and regulatory filings (Loughran and McDonald 2011; Li 2006; Price et al. 2012; Mayew and Venkatachalam 2011; Solomon 2012), the investigation of textual analysis to infer intangible information for firms (such as reputation, culture and corporate social responsibility) remains a gap in the literature.

A primary contribution of this thesis is the introduction of topic modeling to quantify intangible information for publicly listed companies. The traditional approach to inferring intangible information has relied upon survey-based methods (see Edmans 2011). In this thesis we argue that while surveys often provide deep insights into the perceptions of a firm's stakeholders, they are manual and time-consuming to produce, and are thus limited in scope with regards to the number of questions they can ask, the number of companies they can cover and their timeliness to collect and process responses. By contrast, automated measures seek to infer stakeholders' perceptions at a higher frequency and for a large number of firms, providing a significant advantage over current research practice (Popadak 2013).

Second, we contribute to NLP literature by developing general methods to incorporate financial domain knowledge into topic models. Prior NLP studies have demonstrated that the results from LDA are mixed when applied to unconventional data sets (Hong and Davison 2010; Zhao et al. 2011). This is often because there are too few documents, the documents are too short, or contain many topics (Tang et al. 2014). Consequently, a purely unsupervised topic model may recover topics which represent strong statistical patterns but do not correspond to user expectations of semantically meaningful topics. In this thesis, we combine financial domain knowledge with statistical learning to design applications for financial analysis. We demonstrate how to steer a LDA model towards topics of interest based upon investors' modeling goals.

1.5.2 Empirical analysis of intangible information

Our findings contribute to the growing body of evidence documenting investors' underreaction to intangible information. The high costs associated with gathering and processing unstructured data suggests that intangible information may be overlooked by investors compared to readily-accessible and structured financial data (see Da et al. 2011). In this thesis, we investigate the relation between measures of intangible information and financial analysts' "errors-in-expectations" of firms' earnings. If proxies for intangible assets (liabilities) cause positive (negative) stock returns because of financial analysts' "errors-in-expectations", then financial analysts' forecasts of future earnings should be systematically too low (high) relative to actual earnings.

1.6 Thesis overview

This thesis is interdisciplinary in nature and is broadly divided into essays written from an Information Retrieval/NLP perspective (Chapters 2-4) and essays written from an empirical finance and organizational research perspective (Chapters 5-7). The first three essays are methodological and are intended to highlight the merits of probabilistic topic models to infer intangible information. The second three essays investigate the statistical relation between measures of intangible information and firms' earnings. Each chapter in this thesis can be read independently. A brief description of the chapters in this thesis is outlined below:

Chapter 2 describes an automated approach to infer emotions in text. Negative emotions such as anger, contempt and disgust are often linked to specific triggering events. In this study, we investigate emotion-invoking financial media texts and a potential link to investors' subsequent trading decisions.

Chapter 3 describes an automated approach to evaluate the quality of corporate social responsibility (CSR) disclosures. Prior organizational studies suggest that companies publish CSR reports merely for symbolic purposes rather than to provide accountability to investors. Our results may be of interest to investors seeking to integrate environmental sustainability considerations into their investment decisions.

Chapter 4 investigates a textual analysis of central bank communications. Prior studies suggest that transparent central bank communications can help mitigate a financial crisis while ineffective communications may exacerbate one. In this study we design an automated system to predict the impact of central bank communications on investors' interest rate expectations. Our findings contribute to highlighting the role of a central bank's reputation building activities to gain the credibility and confidence of investors.

Chapter 5 conducts a textual analysis of media news for a sample of Chinese ADRs and compares the topics discussed by English- and Chinese-language media outlets. Our findings suggest that a large proportion of corporate governance news articles published by Chinese media outlets appear to be overlooked by the English media. Next, we investigate investors' attention to Chinese corporate governance news. Our findings are consistent with the notion that high costs associated with searching, translating and processing foreign language news creates informational frictions for foreign investors.

Chapter 6 presents a novel social media dataset and employs an automated computational linguistics technique to infer employees' perceptions of corporate culture. To date, investors' efforts to 'look inside' a company have been hampered by a lack of data. Traditional survey-based measures are manual and time-consuming to produce, limited in scope with regards to the number of questions they can ask, the number of companies they can cover and their timeliness to collect and process responses. This study seeks to overcome these limitations by inferring employees' perceptions from social media. Our study highlights the merits of textual analysis for automated corporate culture analysis and builds on the growing body of evidence which suggests that intangible information is not fully exploited by investors.

Chapter 7 employs a probabilistic topic model to infer journalists' and other stakeholders' attributions of firms' poor CSR practices. The approach seeks to automatically detect contextual information and semantic meaning in text. Attributions range from allegations/criticisms over poor CSR to more material concerns reflecting corporate difficulties and litigation risk. Our findings suggest that journalists' and stakeholders' attributions of material CSR concerns are negatively associated with stock returns and firms' future earnings surprises.

Chapter 8 concludes the thesis, discusses the limitations of unstructured data for financial analysis and provides some final thoughts.

1.7 Origins

The work contributing to this thesis has been peer reviewed:

Chapter 2 is a reprint of the paper: Moniz, A. and de Jong, F., (2014c), Classifying the impact of negative affect expressed by the financial media on investor behavior, Proceedings of the 6th Conference on Information Interaction in Context (IIiX). The paper was separately peer-reviewed and presented at the 2014 Behavioral Finance Working Group Conference on Emotional Finance, University of London.

Chapter 3 is a reprint of the paper: Moniz, A. and de Jong, F., (2015), Analysis of companies' non-financial disclosures: Ontology learning by topic modeling, The Semantic Web: ESWC 2015, Springer 2015 Lecture Notes in Computer Science.

Chapter 4 is a reprint of the paper: Moniz, A. and de Jong, F., (2014b), Predicting the impact of central bank communications on financial market investors' interest rate expectations. The Semantic Web: ESWC 2014, Springer 2014 Lecture Notes in Computer Science. The study received best paper award and was cited in a handbook for text mining published by the Bank of England (Bholat et al. 2015).

Chapter 6 is an extended version of Moniz, A. and de Jong, F., (2014a), Sentiment Analysis and the Impact of Employee Satisfaction on Firm Earnings. Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Springer 2014 Lecture Notes in Computer Science. The paper was presented at the Journal of Accounting and Economics Conference in 2015.

Chapter 7 is an extended version of the paper: Moniz, A. and de Jong, F., (2014d), Reputational DAMAGE: Classifying the impact of allegations of irresponsible corporate behavior expressed in the financial media, 34th International Symposium on Forecasting 2014 conference proceedings.

To reiterate, Chapters 2-4 are short Information Retrieval/NLP papers. Chapters 5-7 are longer, empirical finance papers. Each chapter can be read independently.

1.8 Declaration of contribution

In this section, I declare my contribution to the different chapters of this thesis and also acknowledge the contribution of other parties where relevant. The majority of the work has been undertaken independently by the author of this thesis. For each chapter, the author formulated the research question, performed the literature review, conducted the data analysis, interpreted the findings, and wrote the manuscript. The promotors and the co-promotor provided detailed feedback for each chapter which was duly incorporated by the author.

.

Chapter 2

Classifying the influence of negative affect expressed by the financial media on investor behavior[1]

ABSTRACT Prior text mining studies have documented a causal link between human emotions and stock market patterns yet relatively little research exists into what triggers these emotions. This study aims to bridge the gap by inferring emotions in text. Underlying our approach lies Attribution Theory, which addresses how observers form causal inferences and moral judgments to explain human behavior, particularly those with negative outcomes.

[1] This paper was published in: Moniz, A. and de Jong, F., (2014), Classifying the impact of negative affect expressed by the financial media on investor behavior, Proceedings of the 6th Conference on Information Interaction in Context (IIiX).

2.1 Introduction

There is a large body of work on affective text mining (e.g. product reputation mining, customer opinion extraction, and sentiment classification) (see Bollen et al. 2010), yet relatively little research has explored the social psychological mechanism that links affective terms expressed by the media to human behavior (Lee et al. 2010a; Jin et al. 2011; Kim and Cameron 2011). The stock market provides an interesting setting to evaluate text-based emotion cause detection. Given limited time and cognitive processing abilities, investors often turn to the financial media to determine the salience of news when forming their investment decisions (Deephouse 2000; Tetlock 2007). Prior studies document a link between emotions such as investor fear and happiness to stock market patterns (Bollen et al. 2010; Tetlock 2007; Hirshleifer and Shumway 2003). Such emotions are known as attribution-independent emotions (Lazarus 1991; Choi and Lin 2009) because they lack clear attribution to particular events. The examination of investors' attribution-dependent emotions (e.g. anger, contempt and disgust) presents a gap in the affective text mining literature. Such emotions are linked to specific triggering events and, in the context of the stock market, may be invoked when a corporation is accused of acts of irresponsible behavior. From an applications perspective, our results may be of interest to investors seeking to interpret attributions and emotions expressed by the financial media as part of their investment analysis (see Tetlock 2007, 2008). It can therefore be seen as an illustration of the cross-over potential of text mining.

The rest of this study is structured as follows. Section 2.2 draws on literature from the fields of social psychology, organizational studies and emotion-based textual analysis and discusses the influence of media attributions and emotions on public perceptions with regard to investors' behavior. Section 2.3 describes the components of the proposed joint emotion-topic model that combines a measure of media pessimism and a probabilistic topic model in an ensemble tree. In Section 2.4 we outline our financial media corpus, present the experiments and discuss the results. Section 2.5 concludes.

2.2 Related literature

2.2.1 Background: crisis emotions

In this study we examine the role of negative affect (Watson and Clark 1984), defined as the human experience of negative emotions (e.g. anger, fear, disgust, guilt, and nervousness), on investor behavior. Prior research in the fields of social psychology (Baumeister et al. 2001; Floyd and Voludakis 1999; Pennebaker et al. 2003) and finance (Tetlock et al. 2007) suggest that, due to cognitive limitations, negative information has a greater impact on investor behavior than positive information. Underlying our approach lies Attribution Theory, the dominant theory developed in the field of social psychology (Weiner 1985), which addresses how observers form causal inferences and moral judgments to explain irresponsible behavior. The theory holds that people make judgments about the causes of events, especially unexpected events with negative outcomes.

The financial media play a critical role in influencing the reputation of companies (Deephouse 2000) by expressing views that are often written to provoke a public reaction (Strapparava and Mihalcea 2008). Consider, for example, a news report of a factory fire that causes employee fatalities. If the news coverage emphasizes the firm's intentional negligence (Kim and Cameron 2011), anger might dominate the public's response; the public may consider the firm an object of blame for not controlling the crisis or preventing it from occurring. If the news story focuses on the victims' personal lives or their families' suffering, a feeling of sadness may be invoked (Kim and Cameron 2011). Alternatively, if the media emphasize that the accident may occur again, fear may dominate the public's emotions (Lazarus 1991) which may result in a boycott of the firm's products (Murphy et al. 2009). Consequently, how the media perceive, feel about, and evaluate corporate behavior can influence investors' behavior (Deephouse 2000). To our knowledge, there is no empirical evidence that provides a large-scale test of this proposition. Our study aims to demonstrate how the field of text mining can contribute to the generation of an empirical foundation for this claim, and provide a deeper insight into the kind of individual and collective human behavior exhibited in response to corporate allegations. This is assessed by inferring media attributions of irresponsible behavior during a corporate crisis and evaluating the impact on investors' behavior.

2.2.2 Affective text mining

To our knowledge, existing studies of emotional disposition conducted in the field of finance have provided empirical results at the stock market index level. Using data from twenty-six stock exchanges, Hirshleifer and Shumway (2003) suggest that investors are more likely to be in a good mood on a sunny day and consequently more inclined to buy stocks. More recently, measures of collective mood states (calm, alert, sure, vital, kind, and happy) derived from daily aggregated Twitter feeds have been found to be correlated to the value of the Dow Jones Industrial Average (Bollen et al. 2010). These aggregate studies, however, do not provide an insight into investor behavior as a result of emotion-inducing company specific events. We draw on Natural Language Processing (NLP) literature (Lee et al. 2010a; Alm 2009) to consider a deeper understanding of emotions by assuming that emotions are invoked by the perception of external events that in turn trigger reactions (Chen et al. 2010; Plutchik 1962).

Our proposed approach employs a joint emotion-topic model to mine affective content (Bao et al. 2009; Kan and Ren 2011) and is motivated by research in the field of social psychology which suggests that emotions are formed as mixtures from a limited number of primary emotions (Plutchik 1962, 1980). The approach adopted in this study draws on the analogy:*"[I]t is necessary to conceive of the primary emotions as analogous to hues, which may vary in degree of intermixture (saturation) as well as intensity. The primary emotions vary in degree of similarity to one another, just as do colors"* (Plutchik 1980). The mixture distributions, known as 'dyads', include outrage, a combination of primary emotions surprise and anger, contempt a blend of disgust and anger, and remorse, an amalgam of sadness and disgust. We employ an implementation of Latent Dirichlet Allocation (LDA) (Blei et al. 2003) to model this insight.

2.3 Joint emotion-topic model

In this section we describe the components of the system, provide an evaluation and a discussion of the results. The first phase computes a measure of media pessimism expressed in documents. The second phase implements a topic model to discover contextual information, by inferring negative affect associated with media attributions of corporate culpability. The final stage combines the two components in an ensemble tree.

2.3.1 Media pessimism

Following prior text mining studies in the field of finance (Murphy et al. 2009; Loughran and McDonald 2011), we define a document as a financial media allegation of irresponsible corporate behavior and compute a measure of media pessimism by counting terms using the General Inquirer dictionary (Tetlock et al. 2008). The dictionary classifies words according to multiple categories, including positive, negative and various emotions. The dictionary contains 1,915 positive words and 2,291 negative words. Negative terms include: ‘accident’, ‘error’, ‘negligence’ and ‘disaster’. We perform a pre-processing step that consists of stemming using the Porter2 algorithm and stop word removal before measuring the standardized fraction of negative terms in a document (Tetlock et al. 2008). We include the media pessimism measure as a component within the ensemble tree.

2.3.2 Emotion-topic model

The media pessimism component treats negative terms individually and cannot discover the contextual information within the document to associate attributions of blame. The second phase therefore extends the approach and employs a LDA model (Blei et al. 2003) to infer negative affect (Bao et al. 2009; Kan and Ren 2011). We label this the emotion-topic model.

LDA represents each document as a probability distribution over latent topics, where each topic is modeled by a probability distribution of words. In Titov and McDonald (2008), LDA is found to capture global topics in documents; defining document clusterings into specific types rather than rateable aspects within the individual documents. We utilize this property and apply a LDA model to documents using only negative affect terms from the General Inquirer dictionary. We define negative affect as negative terms within the emotion categories "Pain", "Feel" and "EMOT" of the General Inquirer dictionary. We use then these terms to seed the LDA model (see Jagarlamudi et al. 2012). We implement standard settings for LDA hyper-parameters with $\alpha = 50/K$ and $\beta=.01$ where K is the number of topics (Griffiths and Steyvers 2004), and adopt a heuristic approach to set the number of topics equal to four. This choice is based on prior studies that suggest that four negative emotions (anger, fright, anxiety, and sadness) dominate the public's emotions during times of crisis (Jin et al. 2011). Table 2.1 identifies the top terms associated with the topic clusters. The task of annotating labels to topic clusters is challenging because the English language 'does not contain emotion words for certain combinations' of dyads (Plutchik 1980). We therefore manually annotate labels associated with the inferred topic clusters.

Table 2.1: Emotion topic clusters LDA

This table reports the top five terms for each topic cluster and their associated probabilities inferred using the Latent Dirichlet Allocation (LDA) algorithm (Blei et al. 2003). In LDA, a topic is modeled as a probability distribution over a set of words represented by a vocabulary and a document as a probability distribution over a set of topics. We implement standard settings for LDA hyperparameters with $\alpha = 50/K$ and β=.01 following (Griffiths and Steyvers 2004). Topic labels are manually annotated to aid the reader's interpretation by drawing upon social psychology literature (see Plutchik 1962).

fear	anger	remorse	contempt
nervous	touchy	sorrow	outcry
twitch	concern	bereavement	sufferer
misunderstand	overflow	lone	loveless
helpless	concern	estranged	rot
hysterical	angry	mortify	rage

We include the resulting document topic probabilities as components within the ensemble tree.

2.4 Experiments

In this section we discuss our corpus of financial media news. We then outline the evaluation of the ensemble classification tree, present the results and provide a discussion.

2.4.1 Data

Our news source is a corpus created from Dow Jones Newswires (DJNS); a source considered to influence investor sentiment (Tetlock 2007). News articles are retrieved from financial blogs, (e.g. MarketWatch.com), on-line newspapers (e.g. The Wall Street Journal) and financial magazines (e.g. Barron.com). We include the 'Editorial Commentary' and the 'Letters to the Editor' sections on the assumption that these articles contain more opinionated views than fact based articles (Kozareva et al. 2007). We conduct keyword searches on the headline and the first sentence of news stories that contain the terms 'accusation' or 'allegation' in lemmatized form. These terms are chosen because they convey negative connotations of corporate behavior, though they are insufficient in their own right to determine the nature, severity and cause of an allegation for an investor to determine the potential impact on a firm's stock market patterns (Deephouse 2000). Drawing on prior organizational studies, we search for news related to companies in Fortune magazine's list of the 'World's Most Admired Companies' (see Levering et al. 1984). Prior studies deem this group of firms to be 'newsworthy' of journalists' attention (Deephouse 2000) and more likely to be negatively impacted by allegations (Kozareva et al. 2007). Our corpus consists of 35,678 daily news stories for 598 global companies for the period 1 January 2009 to 31 December 2013.

2.4.2 Experimental setup

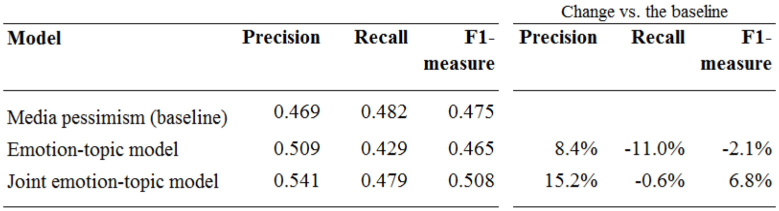
The goal of ensemble methods is to combine the prediction of several models built with a given learning algorithm in order to improve the generalizability and robustness over a single model. We use the Random Forest algorithm (Breiman 2001) to combine the system components and to introduce randomness into the classifier construction.

To classify the likelihood that a given media allegation associated with an act of corporate irresponsible behavior will negatively impact investors' behavior, we compute a measure of investor sentiment obtained from stock market patterns. All else equal, a fall in a company's share price on the day of the announcement implies that investors assess the allegation news to be detrimental to the company's reputation (cf. MacKinlay 1997). Consequently, we define a Boolean which equals one if the change in a company's stock market pattern is negative on the day of the allegation announcement, and zero otherwise. To control for exogenous events that may be announced on the same day as the media allegation we impose a second condition such that the magnitude of the fall in the company's share price must exceed any observed fall in the overall stock market (MSCI All Country World) index. This constraint implies that the fall in a company's share price can be attributed to the allegation news rather than exogenous stock market conditions (cf. Fama 1965).

Experiments were validated using 10-fold cross validation. The dataset is divided into 10 equal sized sets; the classifier is trained on 9 datasets and tested on the remaining dataset. The process is repeated 10 times and we calculate the average across folds. For evaluation, we select precision and recall measures and for completeness include the F1-measure. Table 2.2 displays the evaluation metrics for each of model components and the system.

Table 2.2: Evaluation of the emotion-topic model

This table reports the evaluation metrics of the joint emotion-topic model together with its components. Recall is defined as TP/(TP+FN), precision is measured as TP/(TP+FP) and the F1 measure equals (precision x recall)/(precision + recall). TP refers to the number of true positive classifications, FP refers to the number of false positive classifications, and FN refers to false negative classifications.



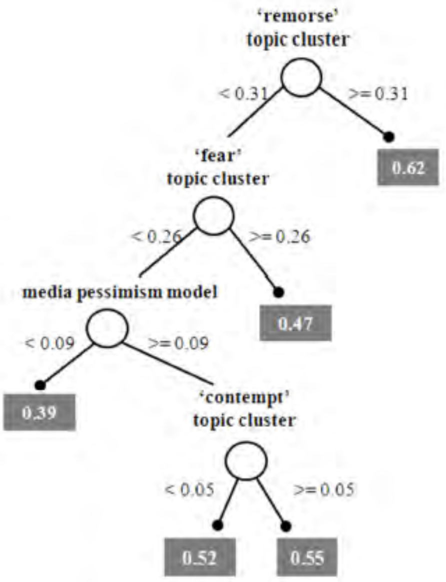
				Change vs. the baseline		
Model	**Precision**	**Recall**	**F1-measure**	**Precision**	**Recall**	**F1-measure**
Media pessimism (baseline)	0.469	0.482	0.475			
Emotion-topic model	0.509	0.429	0.465	8.4%	-11.0%	-2.1%
Joint emotion-topic model	0.541	0.479	0.508	15.2%	-0.6%	6.8%

2.4.3 Discussion

Our findings indicate that the term counting and topic modeling approaches capture distinct, yet complementary dimensions of media sentiment. Precision for the joint emotion-topic model improves by 15% versus the baseline. To aid our understanding of the system, Figure 2.1 displays the decision tree results for one of the folds. The numbers in the grey boxes provide the associated probability values associated with the likelihood of a negative stock market pattern. A value of 1 indicates a 100% likelihood of a negative stock market pattern for the company on the day of the media allegation.

Figure 2.1: Illustrative example of the joint emotion-topic model

This figure provides an example classification tree from one fold of the joint emotion-topic model. Each circle represents an internal node which evaluates a decision function to determine which child node to visit next. The aspects associated with the decision function are labelled above the nodes and the threshold values on each side. The grey boxes represent the terminal nodes in the tree and provide the estimated probabilities associated with a negative stock market pattern, measured as the daily change in a company's share price minus the daily change of the overall stock market (MSCI All Country World). A value of 1 implies there is a 100% likelihood of a negative stock market pattern on the day of the media allegation.



Our findings suggest the presence of a hierarchical relationship between negative affect expressed in the financial media and a company's stock market patterns. The dominance of the emotions of remorse and fear is consistent with prior studies that document investors' risk-averse behavior during crises (Kim and Cameron 2011; Choi and Lin 2009). However, our results also provide new insight into negative affect and investor behavior. In particular, the 'contempt' topic cluster appears to be a more important predictor of negative stock market patterns when related to acts of corporate irresponsible behavior than the emotion of fear.

2.5 Conclusion

In this study, we examine the relationship between acts of corporate irresponsible behavior, the associated negative affect expressed by the financial media and the impact on investors' behavior. Prior studies identify a statistical relationship between investors' emotions and stock market patterns but do not provide a theory to explain this link. Our study aims to demonstrate how the field of text mining can contribute to the generation of an empirical foundation to test theories developed in the fields of social psychological and organizational studies. By comparing a traditional measure of media pessimism, as adopted by prior text mining studies in the field of finance, we find that emotions appear to represent a distinct dimension of media sentiment and contain incremental information for the prediction of a company's stock market patterns.

Chapter 3

Analysis of companies’ non-financial disclosures: Ontology learning by topic modeling[1]

ABSTRACT Prior studies highlight the merits of integrating Linked Data to aid investors’ analyses of company financial disclosures. Non-financial disclosures, including reporting on a company’s environmental footprint (corporate sustainability), remains an unexplored area of research. One reason cited by investors is the need for earth science knowledge to interpret such disclosures. To address this challenge, we propose an automated system which employs Latent Dirichlet Allocation (LDA) for the discovery of earth science topics incorporate sustainability text. The LDA model is seeded with a vocabulary generated by terms retrieved via a SPARQL endpoint. The terms are seeded as lexical priors into the LDA model. An ensemble tree combines the resulting topic probabilities and classifies the quality of sustainability disclosures using domain expert ratings published by Google Finance. From an applications stance, our results may be of interest to investors seeking to integrate corporate sustainability considerations into their investment decisions.

[1] This paper was published in: Moniz, A. and de Jong, F., (2015), Analysis of companies' non-financial disclosures: Ontology learning by topic modeling. The Semantic Web: ESWC 2015, Springer 2015 Lecture Notes in Computer Science.

3.1 Introduction

Prior studies highlight the benefits of employing Linked Data for investment analysis, by combining information from DBpedia, stock market patterns and different taxonomy versions of companies' accounting statements (Kämpgen et al. 2014). Increasingly, investors and regulators are demanding companies to disclose non-financial information, particularly firms' impacts on the environment (referred to as sustainability) (Coburn and Cook 2014). The voluntary nature of corporate sustainability reporting has resulted in the publication of inconsistent and incomplete information (Coburn and Cook 2014). This has inhibited the manual creation of ontologies (O'Riain et al. 2012; Wei et al. 2009). In this study, we employ an automated ontology learning system to overcome this challenge. The proposed system, labelled SPARQL LDA, employs Latent Dirichlet Allocation (LDA) (Blei et al. 2003) for the discovery of topics to represent ontology concepts (Wong et al. 2011; Cimiano 2006; Wei et al. 2009). The system works in three phases. The first phase employs a Naïve Bayesian model to categorize text in sustainability disclosures. The model detects text related to a firm's climate change impacts and aggregates sentences to create a composite document. The second phase employs a LDA topic model to detect contextual information in text. Topics are learned by retrieving terms via a SPARQL endpoint which are seeded as lexical priors into the LDA model. The final phase combines the LDA topic probabilities in an ensemble model and classifies the quality of corporate sustainability reporting using publicly available disclosure ratings.

The rest of this study is structured as follows: Section 3.2 provides a brief overview of relevant sustainability datasets. In Section 3.3 we develop a system to evaluate the quality of corporates' sustainability disclosures. Section 3.4 provides an empirical evaluation of the proposed system. We conclude in Section 3.5.

3.2 Environmental sustainability datasets

Prior earth science literature has explored the benefits of incorporating Semantic Web technologies to predict the impacts of climate change (Pouchard et al. 2013; Bozic et al. 2014; Emile-Geay 2013; Tilmes et al. 2013). To our knowledge, literature has not considered the implications for companies or government regulatory policy. To aid such analysis we highlight two publicly available datasets. The US Global Change Research Act of 1990 requires a National Climate Assessment (NCA) report (Melillo et al. 2014) on the impact of climate change and affected industries. This includes a Global Change Information System (GCIS) which stores climate change metadata. GCIS resources are exported into a triple store queryable through a public SPARQL interface. A second dataset, published under the "Key stats and ratios" section of Google Finance, provides ratings to evaluate the quality of firms' sustainability disclosures. These ratings are collected by the Carbon Disclosure Project (CDP), an initiative led by the United Nations, and are computed from annual surveys of domain experts. The highest CDP rating, 'A', corresponds to companies that are perceived to have published comprehensive climate change disclosures. The lowest rating, "E", corresponds to companies with poor quality disclosures.

3.3 Model of corporate sustainability

3.3.1 Climate change aspect detection

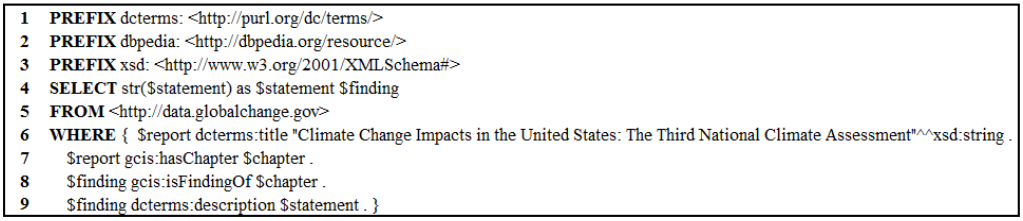
The first phase of the system employs a Naïve Bayesian classifier to detect salient aspects in text. A pre-processing step selects classification features from Wikipedia’s ‘Carbon emissions reporting’ page. The page provides an overview of corporate environmental reporting issues. We select the 10 most frequently occurring unigrams and bigrams as classification features: “climate”, “climate change”, “emissions”, “emitters”, “gas”, “ghg”, “greenhouse”, “scope 1”, “scope 2”, “scope 3”.

3.3.2 LDA topic model

The second phase of the system employs a LDA model (Blei et al. 2003) for the discovery of topics represented as ontology concepts (Wong et al. 2011; Cimiano 2006; Wei et al. 2009; Zavitsanos et al. 2007). In LDA, a topic is modeled as a probability distribution over a set of words represented by a vocabulary and a document as a probability distribution over a set of topics. Our approach departs from a traditional LDA model (Blei et al. 2003) by seeding terms as lexical priors following the approach of Jagarlamudi et al. (2012). Figure 3.1 displays the SPARQL query which retrieves the key recommendations from the latest NCA report using the GCIS interface (see Section 3.2). The unique terms (excluding stopwords) generated by the query form the LDA model’s vocabulary.

Figure 3.1: SPARQL query to retrieve earth science terms

This figure displays the SPARQL query used to retrieve key recommendations from the GCIS interface.



```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT str($statement) as $statement $finding
FROM <http://data.globalchange.gov>
WHERE {  $report dcterms:title "Climate Change Impacts in the United States: The Third National Climate Assessment"^^xsd:string .
   $report gcis:hasChapter $chapter .
   $finding gcis:isFindingOf $chapter .
   $finding dcterms:description $statement . }
```

We implement standard settings for LDA hyperparameters with α = 50/K and β=.01 (Griffiths and Steyvers 2004). The number of topics, K, is set to five following a heuristic approach based on the number of climate change topics reported in the latest NCA report (see Melillo et al. 2014). Table 3.1 displays the top terms associated with the topic clusters. Cluster labels are manually annotated to aid the reader’s interpretation.

Table 3.1: Environmental sustainability topic clusters

This table reports the top terms for each topic cluster and their associated probabilities inferred using the Latent Dirichlet Allocation (LDA) algorithm (Blei et al. 2003). In LDA, a topic is modeled as a probability distribution over a set of words represented by a vocabulary and a document as a probability distribution over a set of topics. We implement standard settings for LDA hyperparameters with α = 50/K and β=.01 following (Griffiths and Steyvers 2004). Topic labels are manually annotated to aid the reader’s interpretation.

footprint	mitigation	adaptation	monitoring	risks
emissions	processes	systems	monitoring	risks
impacts	responses	adaptation	usage	regulatory
ocean	plans	goal	volume	reporting
climates	requirements	thresholds	percentile	policymakers
ecosystems	reported	technology	stabilizing	trends
society	estimates	operational	target	economic
reef	measures	achieving	consumption	shifts
glacier	mitigation	improvements	percent	effects
forest	research	target	capacity	changing

The outcome of the model is a finer-grained categorization of companies’ disclosures based on topics discussed by the online scientific community. The probabilities associated with each topic cluster are included as components within the ensemble tree.

3.4 Ensemble model

In this section we outline the evaluation of the ensemble classification tree, present the results and briefly conclude.

3.4.1 Data

Sustainability disclosures are reported annually on company websites. We retrieve a sample of 443 reports via the Google search query: "sustainability report type:pdf site:" followed by companies' urls obtained from DBpedia (dbpedia-owl:wikiPageExternalLink). Document texts are extracted using PDFMiner. To evaluate the ensemble tree's classifications, we create a Boolean which takes a value of one if a company's CDP disclosure is 'A' rated and zero otherwise (see Section 3.2).

3.4.2 Experimental setup

We design the evaluation by comparing two systems. The benchmark employs a traditional LDA model and infers topics using only the underlying collection of documents. The SPARQL LDA system incorporates lexical priors by seeding the SPARQL generated vocabulary (see Section 3.2). Any differences in classification between the two systems can be explained by the different approaches to topic learning. Experiments were validated using 10-fold cross validation. The performance is evaluated in terms of Precision, Recall, and F1-measure. The evaluation metrics are shown in Table 3.2.

Table 3.2: Evaluation of the SPARQL LDA system

This table reports the evaluation metrics of the SPARQL LDA system and the benchmark. The benchmark employs a traditional LDA model and infers topics using only the underlying collection of documents. The SPARQL LDA system incorporates lexical priors (following Jagarlamudi et al. 2012) Seed words are generated from the SPARQL query. Recall is defined as TP/(TP+FN), precision is measured as TP/(TP+FP) and the F1 measure equals (precision x recall)/(precision + recall). TP refers to the number of true positive classifications, FP refers to the number of false positive classifications, and FN refers to false negative classifications. The final row of the table reports the percentage difference between the evaluation metrics of the benchmark and SPARQL LDA system.

System	Precision	Recall	F1-measure
Benchmark	0.52	0.59	0.55
SPARQL LDA	0.69	0.65	0.67
% difference	32.7%	10.2%	21.8%

Precision for the SPARQL LDA system improves by 33% versus the traditional LDA approach.

3.5 Conclusion

The manual building of ontologies is a time-consuming and costly process particularly in fast evolving domains of knowledge such as earth science, where information is updated often. In this study we employ a fully-automated method for learning ontologies to alleviate the need for manual approaches. Our findings point to the benefits of integrating Linked Data for investors' analyses of both financial and non-financial disclosures.

Chapter 4

Predicting the impact of central bank communications on financial market investors' interest rate expectations[1]

ABSTRACT Prior studies suggest that transparent central bank communications can help mitigate a financial crisis while ineffective communications may exacerbate one. In this study we employ a textual analysis of central bank minutes and design an automated system to predict the impact of central bank communications on investors' interest rate expectations. Our findings contribute to highlighting the role of a central bank's reputation building activities to gain the credibility and confidence of investors.

[1] This is a modified version of the paper: Moniz, A. and de Jong, F, (2014), Predicting the impact of central bank communications on financial market investors' interest rate expectations. The Semantic Web: ESWC 2014. Springer 2014 Lecture Notes in Computer Science. The paper received best paper award at the Workshop on Finance and Economics on the Semantic Web (ESWC 2014) and was cited in a handbook for text mining published by the Bank of England (Bholat et al. 2015).

4.1 Introduction

Post the global financial crisis, there has been a dramatic change in the use of central bank communications as a central bank policy instrument (Vayid 2013). Central banks communicate to the financial markets through statements, minutes, speeches, and published reports (Boukus and Rosenberg 2006). Communication is an important tool that a central bank can use to avert a crisis, by providing investors with its assessment of the risks and the measures it views as necessary to reduce those risks within the economy (Meyersson and Karlberg 2012). Previous studies suggest that effective central bank communications can mitigate and potentially prevent a financial crisis; ineffective communications may exacerbate one (Vayid 2013; Viñals 2010). For example, the Swedish central bank, the Riksbank, was criticized because its communications were "not clear or strong enough" leading up to the global financial crisis, such that the bank's information went "unnoticed" (Meyersson and Karlberg 2012; Vayid 2013). In this study, we design an automated system that predicts the impact of central bank communications on interest rate expectations as derived via financial market patterns. For the purposes of this study, we analyze economic sentiment inferred from the Bank of England's 'Monetary Policy Committee Minutes'. The minutes are published each month and discuss the Bank of England's interest rate decisions.

Financial markets scrutinize central bank communications for "*clues and shades of meaning about its assessment of the economy and the direction of where economic policy may be heading*" (Vayid 2013). As a prediction task, the measurement and evaluation of sentiment is challenging due to the complexities and subtleties of interpreting bank communications (Vayid 2013). The formation of economic policy is a balancing act between achieving high economic growth and financial stability, while targeting low inflation (Bank of England 2013). The relative importance of these objectives is dynamic and varies depending on prevailing economic conditions. For example under benign economic conditions, high inflation may be construed by financial market investors as a negative signal for the direction of future interest rates. During the financial crisis of 2007-2009, high inflation was considered to be a positive signal by effectively lowering real interest rates[2] (Danthine 2013). This motivates a need for fine-grained sentiment analysis to automatically detect economic aspects and predict central bank sentiment expressed towards these aspects (Titov and McDonald 2008). Such an approach would provide investors with an automated system to decipher the complexities and interactions of economic aspects, to interpret the consequences of these interactions for the future path of interest rates, and to incorporate the information into their investment decisions. For a central bank, such a system would provide it with the ability to predict the impact of its economic policies on the

[2] The real interest rate is the rate of interest a borrower expects to pay on debt after allowing for inflation and is equal to the nominal interest rate (set by the central bank) minus the rate of inflation (Bank of England 2013).

financial markets. The resulting 'price discovery' process may promote a more efficient functioning of financial markets (Bank of England 2013).

Our approach consists of four phases. First, the system detects salient references to economic aspects associated with economic growth, prices, interest rates and bank lending and employs a multinomial Naive Bayesian model to classify sentences within central bank documents. Economic aspects are identified in a pre-processing step using a link analysis based upon the TextRank algorithm (Mihalcea and Tarau 2004; Brin and Page 1998) and applied to background knowledge obtained from Wikipedia. The second phase measures sentiment associated with each economic aspect, computed by counting terms from the General Inquirer dictionary (Stone et al. 1966). The third phase employs Latent Dirichlet Allocation (LDA) to infer intensifiers/diminishers that may change the meaning of the economic aspects and economic sentiment (Blei et al. 2003; Titov and McDonald 2008). Specifically, the model categorizes whether the magnitude of the economic aspects has 'intensified' or 'diminished' over time (see Kennedy and Inkpen 2011; Polanyi and Zaenen 2004). We refer to the resulting topic clusters as *directional topic clusters*. Finally, an ensemble tree combines the model components to predict the impact of the communications on financial market interest rates over the following day.

The rest of this paper is structured as follows. Section 4.2 draws on literature from the field of macroeconomics and discusses the implications for sentiment analysis and keyword detection. Section 4.3 models the individual components of the system. Section 4.4 outlines the corpus of central bank communications, provides an evaluation of the model components and then discusses the results. Section 4.5 concludes and suggests avenues for future research.

4.2 Related work

4.2.1 Background: central bank research

Post the financial crisis, several central banks have identified communications, particularly 'enhanced forward guidance', as an important policy instrument within their economic toolkit (Vayid 2013, Bank of England 2013). Effective communications enhance a central bank's public transparency, accountability and credibility (Carney 2013), which in turn aids its ability to implement economic policies (Fay and Gravelle 2010). To date, there has been little research into text mining of central bank communications. Fay and Gravelle (2010) analyse the impact of different types of communications (press releases, speeches, interviews, and news conferences) to determine which media sources impact interest rate expectations. The analysis does not, however, classify the language used in the documents. In Boukus and Rosenberg (2006), a term counting approach is adopted to analyze the sentiment contained within the meeting minutes of the US central bank (the Federal Reserve). In Hendry and Madeley (2010), Latent Semantic Analysis is employed to analyze the sentiment contained within the Bank of Canada's minutes. The intention of this study is to design a fine-grained sentiment analysis approach to analyze the impact of central bank communications on financial market investors. To our knowledge, this remains an unexplored avenue of research.

4.2.2 Background sentiment analysis

Traditionally, fine-grained sentiment analysis has been researched for the classification of online user reviews of products and movies (Pang et al. 2002). Readers are often not only interested in the general sentiment towards an aspect but also a detailed opinion analysis for each of these aspects (Titov and McDonald 2008). Evaluation is conducted by comparing model classifications versus ratings provided by users. The evaluation of economic sentiment is arguably a harder task, due to the lack of a clearly defined outcome to assess model performance. For example, which economic variable should be used to evaluate a model's predictions? The relative importance of the aspects (e.g. economic growth/inflation/interest rates) is subjective, may vary over time, and the measurement of the aspects is only known with significant time delay.

The traditional approach to text-mining within the field of finance is to count terms using the General Inquirer dictionary (Tetlock et al. 2008). The dictionary classifies words according to multiple categories, including 1,915 positive words and 2,291 negative words. The General Inquirer was developed for psychology and sociology research. While it is used for text mining, little research has been conducted as to its suitability within finance (Loughran and McDonald 2011). Aspects that are frequently mentioned in central bank communications, such as the terms 'employment', 'unemployment' and 'growth', are not classified by the General Inquirer dictionary. Adjectives are often needed before investors can interpret the patterns in the economy to form their interest rate expectations (Boukus and Rosenberg 2006). Furthermore, the terms 'inflation' and 'low' are classified as negative by the dictionary, yet 'low inflation' is a positive characteristic and indeed achieving this is a central bank's core objective (Bank of England 2013). The terms 'fall' and 'decline' are classified as negative terms in the General Inquirer dictionary, yet the opposite terms 'rise' and 'increase' are not classified at all.

4.2.3 Background: keyword detection

Graph-based algorithms have received much attention (Mihalcea and Tarau 2004) as an approach to keyphrase extraction and are considered to be state-of-the-art unsupervised methods (Liu et al. 2009).

In a graph representation of a document, nodes are words or phrases, and edges represent co-occurrence or semantic relations. The underlying assumption is that all words in the text have some relationship to all other words in the text. Such an approach is statistical, because it links all co-occurring terms without considering their meaning or function in text. Centrality is often used to estimate the importance of a word in a document (Opsahl et al. 2010), and is a way of deciding on the importance of a vertex within a graph that takes into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information (Boudin 2013). The main advantage of such a representation is that selected terms are independent of their language (Litvak and Last 2008).

4.3 Model to predict changes in investors' expectations

In this section we describe the four phases of the system. First, the system detects salient references to economic aspects and employs a multinomial Naive Bayesian model to classify sentences within documents. The second phase measures sentiment expressed for the economic aspects, using a count of terms from the General Inquirer dictionary. The third phase employs a LDA model and categorizes whether the magnitude of the economic aspects has 'intensified' or 'diminished' (Kennedy and Inkpen 2006; Polanyi and Zaenen 2004). Finally, an ensemble tree combines the model components to predict the impact of the communications on financial market interest rates over the following day.

4.3.1 Aspect detection

In Boukus and Rosenberg (2006) it is shown that a term frequency–inverse document frequency (tf-idf) weighting scheme selects infrequent terms that relate to major news events or economic shocks. The weighting scheme, which seeks to scale down frequently occurring terms and scale up rare terms, is commonly used in IR research (see Manning et al. 2008):

$$w_{i,j} = \begin{cases} \frac{\left(1+\log\left(tf_{i,j}\right)\right)}{\left(1+\log\left(a_j\right)\right)} log \frac{N}{df_i} & if\ tf_{i,j} \geq 1 \\ 0 & otherwise \end{cases} \tag{4.1}$$

where N represents the total number of documents in the sample, df_i the number of documents containing at least one occurrence of the i^{th} word, $tf_{i,j}$ the raw count of the i^{th} word in the j^{th} document, and a_j the average word count in the document.

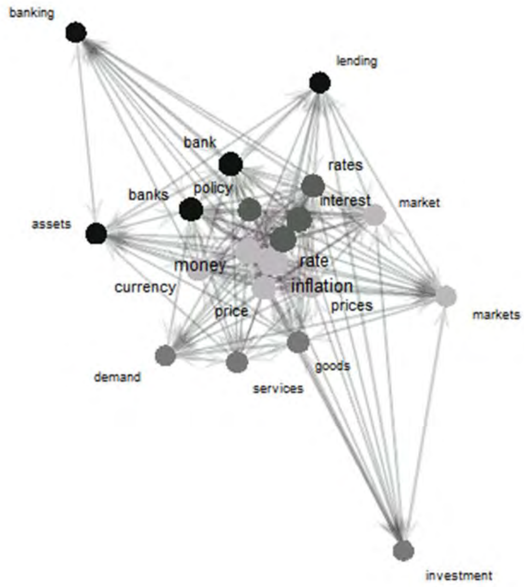
By contrast, our approach is intended to detect the common economic themes that are discussed in central bank communications and are more likely to influence investors' interest rate expectations on a daily basis (see Bank of England 2013). To determine salient references, we employ a link analysis approach that detects the most frequently mentioned terms in Wikipedia's pages on Central Banking and Inflation. TextRank (Mihalcea and Tarau 2004), a ranking algorithm based on the concept of eigenvector centrality, is employed to compute the importance of the nodes in the graph. Each vertex corresponds to a word. A weight, w_{ij}, is assigned to the edge connecting the two vertices, v_i and v_j. The goal is to compute the score of each vertex, which reflects its importance, and use the word types that correspond to the highest scored vertices to form keywords for the text (Boudin 2013). The score for v_i, $S(v_i)$, is initialized with a default value and is computed in an iterative manner until convergence using recursive formula shown in Equation (4.2).

$$S(v_i) = (1-d) + d \; x \sum_{v_j \in Adj(v_i)} \frac{w_{ji}}{\sum_{v_k \in Adj(v_j)} w_{jk}} S(v_j) \tag{4.2}$$

where $Adj(v_i)$ denotes v_i's neighbors and d is the damping factor set to 0.85 (Mihalcea and Tarau 2004). Figure 4.1 displays the resulting clustering of terms. The size of each node is directly proportional to the TextRank score of the respective economic aspect.

Figure 4.1: Link analysis of frequently occurring macroeconomic terms

This figure displays the results of a link analysis based on detecting the most frequently mentioned terms within two Wikipedia pages on Central Banking and Inflation. TextRank (Mihalcea and Tarau 2004), a ranking algorithm based upon the concept of eigenvector centrality, computes the realtive importance of nodes in the graph. Each node corresponds to one of the frequently mentioned terms detected in the two Wikipedia pages. The size of each node is directly proportional to the TextRank score of the respective economic aspect. Different nodes colors reflect different communities identified using the Clauset-Newman-Moore algorithm.



A greedy algorithm is employed to detect communities of terms within the network (see Clauset et al. 2004). The algorithm detects four communities which we label as economic aspects. The *economic growth aspect* detects the frequency of the terms: *'demand'*, *'goods', 'services', 'investment'*. The *prices aspect* detects the terms: *'inflation', 'prices' , 'money', 'markets', 'currency'*. The *interest rate aspect* detects the occurrence of: *'interest', 'rates', 'policy'* and a *bank lending aspect* detects the terms: *'banks', 'lending'* and *'assets'*. It is not surprising to see these terms appear in the link analysis given a central bank's remit is to maintain price and financial stability. The choice of terms is consistent with the text mining research of Boukus and Rosenberg (2006) which identifies 'growth', 'price', 'rate', and 'econom' as the most frequently occurring terms for US central bank communications. Using the four economic aspects, the system next employs a multinomial Naive Bayesian model (McCallum and Nigam 1998) to categorize sentences within each document. The resulting categorization labels form the basis upon which fine-grained sentiment analysis is applied.

4.3.2 Polarity detection

In the second phase, the model computes a measure of economic sentiment associated with each of the four economic aspects. We measure polarity by counting the number of positive (P) versus negative (N) terms, (P−N)/(P+N) identified using the General Inquirer dictionary. In line with Pang et al. (2002), our goal is not to show that a term counting method can perform as well as a machine learning method, but to provide a baseline methodology to measure central bank sentiment and to draw attention to the limitations of the approach that is widely adopted by text mining studies in the field of finance as indicated in Section 4.2. The sentiment metrics that are associated with the economic aspects: economic growth, prices, interest rate and bank lending are labelled $Tone_{growth}$, $Tone_{prices}$, $Tone_{interest\_rates}$ and $Tone_{bank\_lending}$ respectively. A fifth sentiment metric, $Tone_{overall}$, is computed to measure the polarity associated with the overall document, without conditioning upon the *economic aspects*. The five sentiment metrics are included as separate components within the ensemble tree.

4.3.3 Detection of LDA directional topic clusters

Next we extend the baseline term counting method by taking intensifiers and diminishers into account (Kennedy and Inkpen 2006; Polanyi and Zaenen 2004). These are terms that change the degree of the expressed sentiment in a document (see Section 4.2). In the case of central bank communications, the terms describe how economic aspects have changed over time. We employ an implementation of LDA (Blei et al. 2003), and represent each document as a probability distribution over latent topics, where each topic is modeled by a probability distribution of words. In Titov and McDonald (2008), LDA is found to capture the global topics in documents, to the extent that topics do not represent ratable aspects associated with individual documents, but define clusterings of the documents into specific types. For the purposes of training the LDA model, we consider each sentence within each central bank communication to be a separate document. This increases the sample size of the dataset (see Section 4.1) and is intended to improve the robustness of the LDA model for statistical inference. We implement standard settings for LDA hyper-parameters, where $\alpha = 50/K$ and β=.01 (following Griffiths and Steyvers 2004). The number of topics, K, is inferred by maximizing the likelihood of fitting the LDA model over the corpus of documents. We manually annotate two of the topic clusters that capture 'directional' information (Vayid 2013) and appear to act as intensifiers/diminishers of meaning. We label the clusters *directional topic clusters*. Table 4.1 identifies the top terms associated with the two clusters. Representative words are the highest probability document terms for each topic cluster.

Table 4.1 Macroeconomic directional topic clusters

This table reports the top five terms for each topic cluster and their associated probabilities inferred using the Latent Dirichlet Allocation (LDA) algorithm (Blei et al. 2003). In LDA, a topic is modeled as a probability distribution over a set of words represented by a vocabulary and a document as a probability distribution over a set of topics. We implement standard settings for LDA hyperparameters with $\alpha = 50/K$ and β=.01 following (Griffiths and Steyvers 2004). The number of topics, K, is inferred by maximizing the likelihood of fitting the LDA model over the corpus of documents. Topic labels are manually annotated to aid the reader's interpretation based on a heuristic approach. We infer one cluster, labelled 'intensifier' to reflect increasing economic activity and another cluster, 'diminisher' to reflect decreasing economic activity.

'intensifier cluster'		*'diminisher cluster'*	
word	**prob.**	**word**	**prob.**
increase	0.150	moderated	0.190
strong	0.107	lower	0.161
accelerate	0.081	downwards	0.123
strength	0.063	difficult	0.102
support	0.058	less	0.070

Next for each central bank communication the LDA model infers the probabilities associated with the 'intensifier' and 'diminisher' clusters within each of the four economic aspects detected by the Naïve Bayesian classifier. The output of the model is a vector of eight topic probabilities that proxy the central bank's assessment that the economic aspects are intensifying/diminishing. We label the model directional LDA model and the respective probability vectors: $\text{Topic}_{\text{growth}\uparrow}$, $\text{Topic}_{\text{prices}\uparrow}$, $\text{Topic}_{\text{interest\_rates}\uparrow}$ and $\text{Topic}_{\text{bank\_lending}\uparrow}$ if the economic aspects are increasing and $\text{Topic}_{\text{growth}\downarrow}$, $\text{Topic}_{\text{prices}\downarrow}$, $\text{Topic}_{\text{interest\_rates}\downarrow}$ and $\text{Topic}_{\text{bank\_lending}\downarrow}$ if the economic aspects are decreasing. We include the topic probabilities as components within the ensemble tree.

4.4 Experiments

In this section we discuss the corpus of central bank communications and describe the investor patterns data used to evaluate the impact of the central bank communications on investors' interest rate expectations. We then outline the evaluation of the ensemble classification tree, present the results and provide a discussion.

4.4.1 Data

We choose to analyze the interest rate minutes of the Bank of England. As cited in Boukus and Rosenberg (2013), central bank minutes are closely watched by investors to gauge the future direction of economic policies. The Bank of England announces the level of UK interest rates on the first Thursday of every month. The details that underpin this decision are only provided two weeks later and are published in the Bank of England's *'Monetary Policy Committee Minutes'*. The communications are interesting to analyze because changes in investors' expectations on the day of the central bank communication may be attributed to the qualitative information contained within the meeting minutes rather than the interest rate decision announced two weeks before. Minutes typically include summaries of committee members' views on economic conditions and discuss the rationale for their interest rate decisions (Danker and Luecke 2005). The central bank's minutes are, on average, 12 pages long (including a header page), and contain around 55 bullet points, typically with 5 sentences in each bullet. The documents are available from 1997, the year when Parliament voted to give the Bank of England operational independence from the UK government. We retrieve all meeting minutes available between July 1997-March 2014[3] to create a corpus that consists of 199 documents. For the purposes of aspect detection and to train the LDA model, we remove the header page and define a document as an individual sentence within each of the meeting minutes. This expands the corpus to a collection of 53,195 documents.

To evaluate the ensemble tree's predictions we utilize information obtained from financial market patterns. Interest rate futures contracts are financial instruments that enable investors to insure against or speculate on uncertainty about the future level of interest rates (Clews et al. 2000). Changes in the price of the futures contracts therefore reflect changes in investors' views on the future direction in central bank interest rates. Investors' interest rate expectations for the following three, six and twelve months are derived and published daily by the Bank of England. We utilize investors' twelve month ahead forecasts. This data series has the greatest data coverage compared to the three and six month series. Furthermore, the twelve month forecast horizon is consistent with the time horizon over which that the Bank of England conducts its economic policies (Bank of England 2013). To isolate the effect of the central bank communication on investors' expectations, we compute the percentage change in the interest rate futures contract, as measured from the close of business on the day of the communication announcement until the close of business one day after. This narrow time window helps to minimize the influence on investors' interest rate expectations from other financial market factors that may occur at the same time (MacKinlay 1997).

4.4.2 Experiment setup

We design the evaluation in stages in order to enhance our understanding of the system components. For a baseline, we evaluate the system's predictions by using only the tone of the overall document (see Section 4.2). The approach does not take into account individual economic aspects or diminishers/intensifiers (Kennedy and Inkpen 2006; Polanyi and Zaenen 2004). We label the model *naïve tone*. This approach is consistent with the methodology typically adopted by the extant financial literature (Tetlock et al. 2008). Next we compare the outcomes of an ensemble model which combines

[3] Central bank communications announced in August 1997 were excluded from the analysis because the communication document was not readily available in a machine readable format.

the tone associated with each of the economic aspects: economic growth, prices, interest rates and bank lending (see Section 4.2). We label this the *economic aspects model*. A third model compares the outcomes from an ensemble model which combines the intensifiers/diminishers associated with the four economic aspects (see Section 4.3). We label this the *directional LDA model*. Finally, we combine the components in a single ensemble tree and refer to the system as the *joint aspect-polarity* model.

Learning and prediction is performed using an ensemble tree. The goal of ensemble methods is to combine the predictions of several models built with a given learning algorithm in order to improve generalizability and robustness over a single model. We use the Random Forest algorithm (Breiman 2001) which employs a diverse set of classifiers by introducing randomness into the classifier construction. Experiments were validated using five-fold cross validation in which the dataset is broken into five equal sized sets; the classifier is trained on four datasets and tested on the remaining dataset. The process is repeated five times and we calculate the average across folds. For evaluation, we select Mean Absolute Error (MAE), Root Mean Squared Error and Spearman's rho (ρ). We also examine Spearman's rho since prediction may be considered to be a ranking task. The formulae are displayed in Equation (4.3) below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | O_i - E_i| \quad , \quad RMSE = \left[\frac{1}{n}\sum_{i=1}^{n} | O_i - E_i|\right]^2 \quad , \quad \rho = 1 - \frac{6\ \sum(O_i - E_i)^2}{n(n^2 - 1)} \qquad \textbf{(4.3)}$$

where E_i is the model's predicted value, O_i is the realized value, and n is the number of observations. MAE measures the average magnitude of the forecast errors without considering direction; RMSE penalizes errors and gives a relatively high weight to large errors. A smaller value of MAE or RMSE indicates a more accurate prediction. Spearman's rho is a non-parametric measure of the degree of linear association between the predicted and realized values and is bound between the range -1 to +1 (Maritz 1981). A positive Spearman's rho indicates the model's predictive ability; a negative value indicates a poor model fit.

4.4.3 Experiment results

The evaluation metrics from the model components are shown in Table 4.2.

Table 4.2 Evaluation of the macroeconomic joint polarity model

This table reports the evaluation metrics of the joint polarity model together with its components. MAE measures the average magnitude of the forecast errors; RMSE penalizes errors and gives a relatively high weight to large errors. A smaller value of MAE or RMSE indicates a more accurate prediction. Spearman's rho (p) is a non-parametric measure of the degree of linear association between the predicted and realized values and is bound between the range -1 to +1. The asterisks provide the level of significance where *** indicates that the model's predictions versus forecasts are statistically significant at the 0.1% level.

Model	MAE	RMSE	*ρ*
naive tone	0.022	0.016	-0.187 ***
economic aspects	0.018	0.013	-0.044
directional LDA	0.019	0.014	0.041
joint polarity model	0.015	0.011	0.034

The asterisks provide the level of significance where *** indicates that the model's predictions versus forecasts are statistically, negatively significant at the 0.1% level.

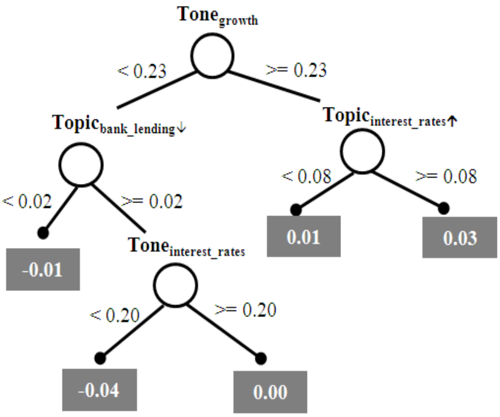
The naïve tone model, which proxies the approach commonly adopted by text mining studies in the field of finance, shows the worst performance. It exhibits the highest MAE and RMSE. The rank correlation of the model's forecasts with realized changes in investors' interest rate expectations is highly statistically negative, implying that documents that are predicted to have a positive/negative impact on investors' interest rate expectations result in the reverse outcome. The economic aspects and directional LDA models exhibit monotonic decreases in MAE and RMSE, suggesting a slight improvement in the model fit. Finally, the joint aspect-polarity model, that includes all model components in the ensemble tree, displays the lowest MAE and RMSE. The mildly positive Spearman's rho is consistent with previous forecasting studies within the field of finance. As cited in Loughran and McDonald (2011), many factors influence the financial markets; a mildly positive correlation is encouraging for the model's predictive power.

4.4.4 Discussion

One interpretation of the experiment results is that multiple aspects are needed to improve the accuracy of the prediction system. The existence of a positive Spearman's rho for the joint model versus a negative Spearman's rho for the naïve tone and economic aspects may be indicative of a non-linear relationship between the components which is only evident when the models are combined rather than considered in isolation. One of the strengths of a regression tree is that it does not assume a functional form, allowing it to detect interactions between model components. To aid our understanding of prediction in the joint model, Figure 4.2 displays the decision tree results for one of the folds. The values in the grey boxes provide the predicted percentage change in investors' interest rate expectations associated with the sentiment contained within the central bank communication. A positive value indicates that the impact is expected to lead to an increase in investors' interest rate expectations, while a negative value indicates an expected decrease in interest rate expectations.

Figure 4.2: Illustrative example of the macroeconomic joint polarity model

This figure provides an example classification tree from one fold of the joint polarity model. Each circle represents an internal node which evaluates a decision function to determine which child node to visit next. The aspects associated with the decision function are labelled above the nodes and the threshold values on each side. The grey boxes represent the terminal nodes in the tree and provide the predicted percentage change in investors' interest rate expectations associated with the sentiment contained within the central bank communication. A positive value indicates that the impact is expected to lead to an increase in investors' interest rate expectations, while a negative value indicates an expected decrease in interest rate expectations.



The regression tree identifies the interaction between the *directional topic clusters* and *Tone* measures. The primary decision in the decision tree is central bank sentiment towards economic growth. The right hand path indicates that if a central bank communication emphasizes positive economic growth and discusses interest rate increases, investors' expectations of future interest rates is predicted to rise by 3%. The left hand path indicates that if a central bank tone towards economic growth is low, discusses declining bank lending and the tone towards interest rates is negative, investors are predicted to lower their expectations of future interest rates changes by 4%.

4.5 Conclusion

The goal of central bank communication is to make messages as clear, simple and understandable as possible to a wide range of audiences (Vayid 2013). In this study, we focus of one specific audience, namely financial market investors. The outcome of our study may feed the design of a system that can predict the impact of central bank communication on the formation of investors' interest rate expectations. The results of the joint aspect-polarity model suggest that investors may benefit by incorporating a measure of central bank sentiment to forecast interest rates. In this study we evaluate model performance using prices from financial market instruments. Post the 2007–09 financial crisis, central banks have broadened the range of their communication, including the use of social media, live broadcasts, podcasts and blogs, to deliver their messages (Vayid 2013). Our study is intended to encourage further research into a wider range of central bank communications including those expressed via social media.

Chapter 5

A multilingual analysis of corporate governance news

ABSTRACT In this study, we conduct a textual analysis of media news for a sample of Chinese ADRs and compare the topics discussed by English- and Chinese-language media outlets. Our findings suggest that a large proportion of corporate governance news articles published by Chinese media outlets appear to be overlooked by the English media. Next, we investigate investors' attention to Chinese corporate governance news. Our findings are consistent with the notion that high costs associated with searching, translating and processing foreign language news creates informational frictions for foreign investors.

纸包不住火*– paper can't wrap up a fire (Chinese proverb, unattributed author)*

5.1 Introduction

In recent years, the U.S. Securities and Exchange Commission (SEC) has raised its concerns about the audit quality of Chinese cross-listed companies and the problems of informational frictions associated with monitoring foreign firms' corporate governance practices (see Ghosh and Wagner 2014). Financial analysts have even coined the phrase "the great transparency wall of China" (Zhu 2009), referring to the hurdles associated investors' abilities to search, translate and process foreign corporate governance news. In this study, we investigate whether informational frictions impede shareholders' abilities to effectively monitor the corporate governance of firms. In theory, one way to mitigate informational frictions is for a foreign firm to 'bond' to a more transparent stock market (e.g. by listing on a U.S. exchange). Bonding provides minority shareholders with greater disclosure, transparency, regulatory and investor scrutiny compared to a firm's local domicile (Coffee 1999, 2002; Stulz 1999; Doidge et al. 2004). To the best of our knowledge, the impact of informational frictions associated with investors' abilities to monitor corporate governance allegations across different languages has yet to be examined by the extant corporate governance literature (see Grinblatt and Keloharju 2001; Domowitz et al. 1998). In this study, we collect 85,067 news articles for a sample of Chinese ADRs over the period 2009-2015. We conduct a textual analysis of English and Chinese language media news, compare the topics discussed by different media outlets and investigate investor attention.

This study makes two important contributions to the literature. First, we contribute to the corporate governance literature by investigating the relation between the transmission mechanism of media allegations of corporate governance misdemeanors and investor attention. While a number of recent studies have conducted event study analyses (Chen et al. 2012; Givoly et al. 2012; Baker et al. 2012; Ang et al. 2012; Darrough et al. 2012), this study aims to provide greater insight into the relation between media coverage of news topics for different media outlets and investors' trading decisions. Second, this study contributes to the growing body of evidence on the existence of informational frictions in equity markets. Prior financial asset pricing literature suggests that investors face constraints and processing frictions particularly for 'soft' non-earnings news due to the intangible nature of such news compared to 'harder' accounting information (see Engelberg 2008; Demers and Vega 2010; Petersen 2004). This study investigates the timing of when investors update their information sets given processing frictions (Ferguson 2015; Hirshleifer and Teoh 2003; DellaVigna and Pollett 2009; Hirshleifer and Teoh 2011). Our study is closely related to Ferguson (2015) but differs in two main respects. First we retrieve a broad sample of publicly available Chinese and English corporate news rather than limit a subset of financial media sources. The breadth of our news corpus means we can test the relation between the salience of different types of news across media outlets and investor attention. Second, we examine investor trading behavior in U.S. listed ADRs, while Ferguson (2015) compares trading behavior in dual listed stocks on U.S. and Chinese exchanges. Our approach is specifically designed to test the reputational bonding hypothesis rather than investors' trading behavior conditional on the type of share listing.

The rest of the study proceeds as follows. Section 5.2 motivates the role of the media as a propagator of legitimacy. Section 5.3 outlines the news corpus and provides descriptive statistics on the English and Chinese media's coverage of news topics. Section 5.4 investigates the conditions under which the English media choose to publish Chinese corporate governance news and documents the main regression results. Section 5.5 concludes this study.

5.2 Theoretical motivation

In theory, a company's decision to engage in a corporate governance misdemeanor (see Dyck et al. 2008; Becker 1968) depends upon whether its management perceives:

$$\begin{aligned}\text{Private benefit to company} &> \text{E(Reputational cost)} + \text{E(Penalty)} \\ &= \sum p_i \text{ x } RC_i \,|\text{i learns of the allegation} + \pi P \qquad \textbf{(5.1)}\end{aligned}$$

where p_i is the probability that stakeholder group i (e.g. consumers, journalists, NGOs, investors, regulators) learns of the misdemeanor, RC_i is the reputational cost associated with the misdemeanor, π is the likelihood of a regulatory penalty and P is the magnitude of such a penalty.

By publishing corporate governance news, the media alters p_i, the probability that a firm's behavior is known to a given stakeholder group. The impact of the news media is greater when the news reaches a larger number of stakeholder groups and is published by a salient and credible media source. The second way in which the media may impact a firm's decision rule is via the perceived reputational cost RC_i of the news. Empirical studies suggest that the media influences stakeholders' information sets by setting the agenda regarding coverage of certain news topics in text and their tone which may alter how stakeholders think about a firm (see McCombs and Shaw 1972; Deephouse 2000). In particular, when the media portray a corporate governance misdemeanor to be the result of intentional actions by the company or its employees (e.g. fraud), reputational damage is more severe (e.g. Hennes et al. 2008). Third, the news media may influence the magnitude of a regulatory fine imposed on a firm for its misdemeanor, either by drawing a regulator's attention (π) or by influencing the way the regulator thinks about the issue (P) (see also Bednar 2012; Miller 2006).

5.3 Data and sample construction

We retrieve media news articles from Dow Jones' Factiva database. One drawback of the academic subscription to the database is the limitation that a maximum of 100 articles can be downloaded at any one time. Due to the high costs of collecting Factiva data, we limit our sample to a set of firms for which informational frictions are likely to be an important issue and retrieve news articles for Chinese ADRs. We rely upon the country of domicile field reported in the CRSP database to identify relevant US listings. Following Ferguson (2015), we select the U.S. market for several reasons. First, the U.S. is the largest stock exchange in the world by market capitalization, increasing the potential sample of Chinese companies with foreign listings. Second, potentially stronger governance practices in the U.S. may reduce the likelihood of informed trading which would otherwise hinder information discovery. Third, the U.S. is one of the last markets to open during the trading day allowing U.S. investors to react to Chinese news releases on the same trading day.

We retrieve news articles written in both traditional Chinese and simplified Chinese for the period 2009-2015. We select this time period for two reasons. First, we choose to exclude the global financial crisis time period when macroeconomic news may have influenced the sensitivity of emerging market stocks to news, potentially biasing our findings (Ferguson 2015). Second, the SEC introduced disclosure regulation which led to a significant increase in cross-listings post 2008 (see Iliev et al. 2014, Ghosh and Wagner 2014). Our choice of sample period may be more representative for regulators seeking to draw conclusions for current policy setting. We download all available media news articles from financial newswires, major newspapers, newswires, press releases and blogs. We classify the media outlets, by employing an automated approach (see Appendix I for details). Our categorization is intended to separate media sources by their potential incentives and access to information (Odzik and Sadka 2013) so that we can test a variety of hypotheses related to investor attention (see Hong and Stein 2000; Chan 2003). We categorise media sources into four outlets: *Financial media* comprise of investment focused outlets, *Newswires* comprise of press releases and wire services, *Newspapers* includes daily national and local newspaper outlets and *Other* comprises of blogs and unclassified sources. Table 5.1 lists the top 5 media sources by the number of news items in the sample for each media outlet.

Table 5.1 Descriptive statistics of news coverage by language and media outlet

This table reports the count of news articles in the sample dataset by language and media outlet. For each media outlet, the five most frequently occurring media sources are displayed, together with the associated number of news articles retrieved (News Items) and the percentage they represent (%). The sample dataset consists of English and Chinese news articles for Chinese ADRs retrieved from the Factiva database. Media outlets are classified into four types: Financial media, Newswires, Newspapers and Other (see Appendix I for details of the classification methodology). Sample period: 2009-2015.

Financial media (Chinese language)

Rank	Source	News Items	%
1	Dow Jones & Company, Inc.	4,041	63%
2	Hong Kong Economic Journal Company	647	10%
3	Hong Kong Economic Times Ltd	551	9%
4	Thomson Reuters (Markets) LLC	511	8%
5	China Business Network Co., Ltd	196	3%

Financial media (English language)

Rank	Source	News Items	%
1	Dow Jones & Company, Inc.	5637	66%
2	Thomson Reuters (Markets) LLC	610	7%
3	Investor's Business Daily	242	3%
4	The Financial Times Ltd	233	3%
5	Business Wire	233	3%

Newswires (Chinese language)

Rank	Source	News Items	%
1	PR Newswire Association, Inc.	1705	40%
2	Infotimes Corporation	605	14%
3	Reuters Ltd	441	10%
4	ET Net Ltd	330	8%
5	N.C.N. Ltd of Xinhua News Agency	320	7%

Newswires (English language)

Rank	Source	News Items	%
1	PR Newswire Association, Inc.	2932	31%
2	GlobeNewswire, Inc.	810	9%
3	Reuters America LLC	677	7%
4	Reuters Ltd	632	7%
5	China Economic Information Service of Xinhua News Agenc	550	6%

Newspapers (Chinese language)

Rank	Source	News Items	%
1	Mingpao.com Limited	565	27%
2	UDN.com	359	17%
3	on.cc (Hong Kong) Ltd	329	15%
4	Beijing News	241	11%
5	Securities Times	221	10%

Newspapers (English language)

Rank	Source	News Items	%
1	South China Morning Post Publishers Ltd	284	21%
2	The New York Times Company	133	10%
3	Australian Associated Press Pty Ltd	65	5%
4	Telegraph Media Group Ltd.	54	4%
5	News UK & Ireland Limited	54	4%

Our findings indicate a concentration of media sources, characteristic of a power law Zipf distribution (Zipf 1932; Breslau et al. 1999).

5.3.1 Classification of news

In this section we describe the approach to classify news. We compute three features to evaluate the salience of news – the topic, relevance for specific companies, and potential financial materiality of the news. News topics are categorised by retrieving metadata from the Factiva database. Metadata are manually assigned by journalists when they write articles. For instance, corporate governance news consists of bribery, corruption, fraud and money laundering news. To determine the relevance of news for specific companies we employ a multinomial Naïve Bayesian text classifier which counts the frequency of company mentions in each article (see Appendix I for details). In a pre-processing step, an automated online translator converts texts into English. Our approach simplifies the need for named entity recognition algorithm. For instance, to recognize that a Chinese article is predominately about Alibaba Inc. requires the ability to detect that 阿里巴巴 is the Chinese name for the company.

The online translator converts all Chinese documents into English. The algorithm then evaluates the relevance of the translated text by counting the frequency of occurrences of company names based upon those reported in the CRSP database. The classifier seeks to distinguish between companies which are the subject of an article versus companies merely referenced in passing. Antweiler and Frank (2004) employ a similar methodology to detect company names in financial media texts. Our approach is computationally efficient and easy to implement (Manning et al. 2008), avoiding the need to manually encode articles which would otherwise require native English and Chinese speakers to read each article. Such an approach would impose substantial time, cost restrictions and potential subjectivity when interpreting news (see Pfarrer et al. 2010). Finally, the financial materiality of news is proxied by authors' references to legal or regulatory consequences in the text. These references are detected using Factiva's metadata. Table 5.2 summarises the coverage of media news for Chinese ADRs by topic, media outlet and language.

Table 5.2: Descriptive statistics of news coverage by topic and media outlet

This table reports the count of news articles in the sample dataset by topic, language and media outlet. Count refers to the number of news articles published by each media outlet for a particular news topic. % indicates the relative number of news articles published on a particular news topic for each media outlet. The sample dataset consists of English and Chinese news articles for Chinese ADRs, retrieved from the Factiva database. Media outlets are classified into four types: Financial media, Newswires, Newspapers and Other (see Appendix I for details of the classification methodology). News articles are classified into topics using metadata tagged to each article in the Factiva database. Panel A reports media coverage of English language news. Panel B reports media coverage of Chinese language news. Sample period: 2009-2015.

Panel A: English language news

	Financial media		Newswires		Newspapers		Other		Total	
News topic	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**
Branding & marketing	1,361	10%	1,355	16%	514	10%	260	9%	3,492	11%
Corporate actions	3,630	26%	1,478	17%	744	14%	552	19%	6,410	21%
CSR	1,274	9%	1,527	18%	1,547	30%	862	29%	5,226	17%
Earnings	6,055	43%	2,651	31%	939	18%	624	21%	10,277	33%
Legal news	627	4%	720	8%	660	13%	320	11%	2,333	8%
Macroeconomic	1,039	7%	859	10%	776	15%	355	12%	3,037	10%
Total	13,986	100%	8,590	100%	5,180	100%	2,973	100%	30,775	100%

Panel B: Chinese language news

	Financial media		Newswires		Newspapers		Other		Total	
News topic	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**	**Count**	**%**
Branding & marketing	3,118	14%	4,101	23%	542	18%	1,989	17%	9,885	18%
Corporate actions	5,259	24%	3,539	20%	551	18%	2,269	20%	11,765	22%
CSR	589	3%	473	3%	186	6%	164	1%	1,472	3%
Earnings	11,100	51%	8,153	46%	1,160	38%	6,204	54%	26,835	49%
Legal news	812	4%	639	4%	451	15%	445	4%	2,395	4%
Macroeconomic	718	3%	680	4%	157	5%	341	3%	1,940	4%
Total	21,596	100%	17,585	100%	3,047	100%	11,412	100%	54,292	100%

Panel A indicates that a large proportion (49%) of news reported by the English language media relates to earnings-related information, while Panel B finds that Chinese language media report a relatively higher proportion of corporate governance news (16% versus 2% for English language media outlets).

5.4 Results and discussion

In this section we investigate the English language media's decision to publish corporate governance news for Chinese companies. Our working premise is that the English media may rely upon the Chinese media to determine the salience of news in the presence of informational frictions. This may be the case if the Chinese news media face lower fixed costs to collect information (potentially due to low search and translation costs) or possess more value-relevant information than the English language media (Ferguson 2015; Feng and Seasholes 2004; Ivkovic and Weisbenner 2005). To investigate media coverage of corporate governance news we employ a logistic regression. The dependent variable, *Media_Cov*, equals one if an English language media outlet publishes corporate governance news for a Chinese company on a given day, and zero otherwise:

$$\begin{aligned} P(\text{Media\_Cov} = 1) = \; & \beta_1 \text{Alleg\_E\_minus1} + \beta_2 \text{Alleg\_C} + \beta_3 \text{Alleg\_C\_minus1} + \beta_4 \text{Violate\_E} + \beta_5 \text{Violate\_C} \\ & + \gamma_1 \text{Finmedia\_E} + \gamma_2 \text{Newspapers \_E} + \gamma_3 \text{Newswires\_E} \\ & + \gamma_4 \text{Finmedia\_C} + \gamma_5 \text{Newspapers \_C} + \gamma_6 \text{Newswires\_C} \\ & + \delta_1 (\text{Violate\_E x Alleg\_E}) + \delta_2 (\text{Violate\_C x Alleg\_C}) \\ & + \delta_3 (\text{Finmedia \_E x Alleg\_E}) + \delta_4 (\text{Newspapers \_E x Alleg\_E}) + \delta_5 (\text{Newswires\_E x Alleg\_E}) \\ & + \delta_6 (\text{Finmedia \_C x Alleg\_C}) + \delta_7 (\text{Newspapers \_C x Alleg\_C}) + \delta_8 (\text{Newswires\_C x Alleg\_C}) \\ & + \text{Controls} \end{aligned} \tag{5.4}$$

The independent variables are a series of indicators. We define *Alleg_E_minus1* equal to one if a Chinese corporate governance news story is published in an English media outlet during the previous day, and zero otherwise. The coefficient β_1 captures the escalation of a Chinese corporate governance news story based upon its prior news coverage in the English media. The indicators *Alleg_C* and *Alleg_C_minus1* are equal to one if corporate governance news is published in a Chinese media outlet during the current and previous calendar day respectively, and zero otherwise. The indicator variables Violate_E and Violate_C equal one if an English or Chinese media source respectively discusses legal/regulatory violations in the text. We further include indicator variables to capture the type of the media outlet that published the news: financial media (*Finmedia*), newswire (*Newswires*) and newspaper (*Newspapers*), where the suffixes *'_E'* and *'_C'* are used to denote English and Chinese media outlets respectively. The remaining variables comprise of interaction terms. To control for the information environment, we include Log(Market Equity), the natural log of the market value of equity, measured as at the end of the preceding calendar year, Num_analysts, the number of sell-side analysts covering a company, and institutional ownership (Ownership). We obtain data on institutional ownership from Thomson Reuters' CDA/Spectrum Institutional Holdings dataset. The metric is equal to the total number of shares owned by institutions as at the end of the prior quarter divided by the number of shares outstanding. If Thompson Reuters does not contain any ownership information for a given stock in a given quarter, we assume that institutional ownership is zero. These controls are intended to proxy for investor sophistication and analysts' coverage of information (Fang and Peress 2009). Finally, we include prior price momentum (Pmom), measured as the signed stock's return measured over the previous 12 months, to control for 'news worthiness' and investor attention (see Da et al. 2011; Carhart 1997). The regression results are reported in Table 5.3.

Table 5.3: Logistic regression of English media coverage of Chinese corporate governance news
This table reports the results of a logistic regression of English language media coverage of Chinese corporate governance news. The dependent variable is an indicator equal to one if an English language media outlet publishes a Chinese corporate governance news article on a given day, and zero otherwise. The sample dataset consists of English and Chinese news articles for Chinese ADRs retrieved from the Factiva database (see text for the descriptions of each variable). Standard errors are clustered by firm following Petersen (2009). For each variable we report corresponding robust t-statistic (in parentheses). Sample period: 2009-2015.

	(1)	(2)	(3)
Alleg_E_minus1	2.272	2.571	2.430
	(3.687)	(4.154)	(3.79)
Alleg_C	0.994	2.606	2.749
	(1.231)	(2.521)	(2.664)
Violate_E_minus1	1.591	1.329	1.369
	(2.999)	(2.388)	(2.462)
Violate_C	3.383	4.292	4.283
	(4.616)	(6.376)	(6.354)
Alleg_C x Violate_C		4.133	4.295
		(3.47)	(3.528)
Alleg_C x Finmedia_C		2.615	2.743
		(2.638)	(2.682)
Alleg_C x Newswires_C		0.746	0.584
		(0.754)	(0.574)
Alleg_C x Newspapers_C		-0.599	-0.892
		(-0.634)	(-0.87)
Alleg_C x Finmedia_C x Violate_C			-0.082
			(-0.057)
Alleg_C x Newswires_C x Violate_C			2.142
			(1.547)
Alleg_C x Newspapers_C x Violate_C			-11.464
			(-0.03)
log(Market Equity)	-0.759	-0.761	-0.761
	(-25.785)	(-25.639)	(-25.623)
Num_analysts	0.047	0.045	0.045
	(2.797)	(2.638)	(2.607)
Ownership	0.373	0.360	0.364
	(3.354)	(3.163)	(3.21)
Pmom	0.030	0.020	0.020
	(0.314)	(0.207)	(0.209)

Column 1 of Table 5.3 indicates that the probability of English language media news coverage increases when the Chinese corporate governance news has already been reported in the English media (i.e. the story has escalated in media attention). Column 2 considers the salience of news by media outlet and severity. The interaction term between *Alleg_C x Violate_C* is statistically significant indicating that English media coverage increases when Chinese journalists attribute legal/regulatory violations. This finding is consistent with the view that the English language media rely upon local (Chinese) media sources to determine the salience of news and "set the agenda" (McCombs and Shaw 1972).

Next we examine the impact of news media coverage for investors' trading decisions. Diamond and Verrecchia (1991) show that more informative disclosures reduce the information advantage of privately informed traders, thereby reducing information asymmetries. If media coverage disseminates new information to a broad group of investors we should observe an increase in share trading volume on the day the article is published (Kim and Verrecchia 1991). In the presence of informational frictions we should only observe a positive relation between media coverage and investors' trading behavior for more salient news sources (see Ferguson 2015). We specify the following regression to investigate the relation between abnormal trading volume (AV) and corporate governance news:

$$\begin{aligned}
\text{AV} = \;& \beta_1\text{Alleg\_E} + \beta_2\text{Alleg\_C} + \beta_3\text{Violate\_E} + \beta_4\text{Violate\_C} \\
& + \gamma_1\text{Finmedia\_E} + \gamma_2\text{Newspapers\_E} + \gamma_3\text{Newswires\_E} \\
& + \gamma_4\text{Finmedia\_C} + \gamma_5\text{Newspapers \_C} + \gamma_6\text{Newswires\_C} \\
& + \delta_3(\text{Finmedia \_E x Alleg\_E}) + \delta_4(\text{Newspapers \_E x Alleg\_E}) + \delta_5(\text{Newswires\_E x Alleg\_E}) \\
& + \delta_6(\text{Finmedia \_C x Alleg\_C}) + \delta_7(\text{Newspapers \_C x Alleg\_C}) + \delta_8(\text{Newswires\_C x Alleg\_C}) \\
& + \text{Controls}
\end{aligned} \qquad \textbf{(5.5)}$$

To measure the impact on trading volumes we follow Frazzini and Lamont (2007) and compute scaled volume (SV), defined as the ratio of share volume for firm j today to firm j's average monthly share volume over the previous 12 months:

$$SV_t^j = \frac{VOL_t^j}{\frac{1}{12}\sum_{s-13}^{-1} VOL_{t+s}^l} \qquad \textbf{(5.6)}$$

Abnormal volume (AV) is defined as scaled volume minus the equal weight average of scaled volume during the month:

$$AV_t^j = SV_t^j - \frac{1}{n}\sum_{j=1}^{n} SV_t^j \qquad \textbf{(5.7)}$$

Under the null hypothesis of no news announcement effect, abnormal trading volumes should on average equal zero. Our set of controls include the prior trading day's abnormal share volume and firm visibility proxied by the number of sell-side analysts covering each company. In addition, we control for industry and year effects. We estimate pooled OLS regressions and we test for the significance of the coefficients using robust standard errors (Petersen 2009). The pooled regression results are reported in Table 5.4.

Table 5.4: Regression of trading volumes and media coverage of corporate governance news

This table reports the OLS estimates of a longitudinal regression of abnormal trading volumes and media coverage of corporate governance news. The dependent variable, AV, measures abnormal volumes and is defined as scaled volume minus the equal weight average of scaled volume during the month, where scaled volume is the ratio of share volume for firm j to firm j's average monthly share volume over the previous 12 months (Frazzini and Lamont 2007 The sample dataset consists of English and Chinese news articles for Chinese ADRs retrieved from the Factiva database. All regressions include control variables for firm characteristics which are excluded for presentation purposes only (see text for details). Standard errors are clustered by firm following Petersen (2009). For each variable we report corresponding robust t-statistic (in parentheses). Sample period: 2009-2015.

	(1)	(2)	(3)
Alleg_E	-0.219	-1.760	-1.577
	(-0.729)	(-2.958)	(-2.473)
Alleg_E_minus1	-0.408	-0.298	-0.292
	(-0.979)	(-0.712)	(-0.696)
Alleg_C	0.614	0.936	0.690
	(1.456)	(1.326)	(0.937)
Alleg_C_minus1	-0.789	-0.793	-0.873
	(-1.676)	(-1.684)	(-1.848)
Violate_E	0.748	0.805	0.466
	(2.916)	(2.103)	(1.683)
Violate_C	-0.081	-0.259	-0.481
	(-0.201)	(-0.624)	(-0.97)
Alleg_E x Finmedia_E		1.156	-0.085
		(1.999)	(-0.115)
Alleg_E x Newswires_E		1.550	1.162
		(2.531)	(1.604)
Alleg_E x Newspapers_E		-0.390	0.912
		(-0.402)	(0.557)
Alleg_C x Finmedia_C		0.824	-0.324
		(1.182)	(-0.242)
Alleg_C x Newswires_C		-0.695	-0.203
		(-0.922)	(-0.179)
Alleg_C x Newspapers_C		-0.737	-0.397
		(-1.009)	(-0.398)
Alleg_E x Finmedia_E x Violate_E			2.478
			(2.287)
Alleg_E x Newswires_E x Violate_E			1.323
			(1.351)
Alleg_E x Newspapers_E x Violate_E			-1.309
			(-0.624)
Alleg_C x Finmedia_C x Violate_C			1.440
			(0.927)
Alleg_C x Newswires_C x Violate_C			0.070
			(0.053)
Alleg_C x Newspapers_C x Violate_C			0.030
			(0.028)

Column 1 suggests that investor attention increases when Chinese corporate governance violations are published by English rather than Chinese language media outlets. Column 2 refines the analysis and suggests that this relation is greater for English language financial media outlets and newswires rather than newspapers or blogs. This is potentially explained by the greater salience and timeliness of information releases for investors' trading decisions (Huberman and Regev 2001; Ferguson 2015). Finally, Column 3 investigates the relation between language, media salience and the severity of corporate governance news. Our findings indicate a statistically significant relation between investor attention and journalists' attributions of legal/regulatory violations when news is published in the English financial media rather than in Chinese media outlets.

Next we evaluate information discovery by measuring the impact of media news on abnormal stock returns (see Griffin et al. 2010). Our main test is an event-study regression specified as:

$$\begin{aligned} \text{FFCAR} = \; & \beta_1\text{Alleg\_E} + \beta_2\text{Alleg\_C} + \beta_3\text{Violate\_E} + \beta_4\text{Violate\_C} \\ & + \gamma_1\text{Finmedia\_E} + \gamma_2\text{Newspapers\_E} + \gamma_3\text{Newswires\_E} \\ & + \gamma_4\text{Finmedia\_C} + \gamma_5\text{Newspapers \_C} + \gamma_6\text{Newswires\_C} \\ & + \delta_3(\text{Finmedia \_E x Alleg\_E}) + \delta_4(\text{Newspapers \_E x Alleg\_E}) + \delta_5(\text{Newswires\_E x Alleg\_E}) \\ & + \delta_6(\text{Finmedia \_C x Alleg\_C}) + \delta_7(\text{Newspapers \_C x Alleg\_C}) + \delta_8(\text{Newswires\_C x Alleg\_C}) \\ & + \text{Controls} \end{aligned} \tag{5.8}$$

The dependent variable, FFCAR, is the firm's abnormal return during the day after the publication of the corporate governance news. We estimate benchmark returns using the Fama-French (1993) three-factor model with an estimation window of [–252,–31] trading days prior to the news announcement. We control for firm size and book-to-market ratios using each firm's log of market capitalization and log of book-to-market equity measured at the end of the calendar year following Fama and French (1992). Table 5.5 reports the regression results.

Table 5.5: Regression of stock returns and media coverage of corporate governance news

This table reports the OLS estimates of a longitudinal regression of abnormal stock returns and media coverage of corporate governance news. The dependent variable, FFCAR, measures daily excess stock returns. We estimate benchmark returns using the Fama-French (1993) three-factor model with an estimation window of [–252,–31] trading days prior to the news announcement. The sample dataset consists of English and Chinese news articles for Chinese ADRs retrieved from the Factiva database. All regressions include control variables for lagged firm returns and other firm characteristics which are excluded for presentation purposes only (see text for details). Standard errors are clustered by firm following Petersen (2009). For each variable we report corresponding robust t-statistic (in parentheses). Sample period: 2009-2015.

	(1)	(2)	(3)
Alleg_E	-0.017	0.016	0.016
	(-3.02)	(1.466)	(1.389)
Alleg_E_minus1	0.008	0.006	0.006
	(1.099)	(0.826)	(0.777)
Alleg_C	0.004	0.004	0.009
	(0.57)	(0.341)	(0.682)
Alleg_C_minus1	0.003	0.003	0.004
	(0.358)	(0.391)	(0.413)
Violate_E	0.002	0.001	0.007
	(0.484)	(0.296)	(1.371)
Violate_C	-0.008	-0.003	-0.006
	(-1.027)	(-0.403)	(-0.696)
Alleg_E x Finmedia_E		-0.032	-0.028
		(-2.821)	(-2.105)
Alleg_E x Newswires_E		-0.024	-0.01
		(-1.286)	(-0.769)
Alleg_E x Newspapers_E		-0.001	0.005
		(-0.045)	(0.157)
Alleg_C x Finmedia_C		-0.026	-0.029
		(-1.537)	(-1.152)
Alleg_C x Newswires_C		0.007	0.008
		(0.511)	(0.387)
Alleg_C x Newspapers_C		0.011	-0.003
		(0.825)	(-0.172)
Alleg_E x Finmedia_E x Violate_E			-0.035
			(-1.969)
Alleg_E x Newswires_E x Violate_E			-0.019
			(-1.021)
Alleg_E x Newspapers_E x Violate_E			-0.018
			(-0.453)
Alleg_C x Finmedia_C x Violate_C			0.004
			(0.153)
Alleg_C x Newswires_C x Violate_C			-0.006
			(-0.24)
Alleg_C x Newspapers_C x Violate_C			0.019
			(0.971)

Column 1 indicates a statistically significant and negative relation between corporate governance news published in English media and subsequent abnormal returns. By contrast, the market reaction to

Chinese media sources is statistically insignificant even after controlling for a potential delayed reaction to news due to differences in time zones. Column 2 includes interaction effects to control for differences in investor reactions across media outlets. Consistent with prior corporate governance literature, the regression results indicate that the negative abnormal return on the day of the news announcement is attributable to investors' reactions to financial media news (see Bushee et al. 2010). Finally, Column 3 finds evidence of a statistically significant and negative relation between the publication of corporate governance violations by English language financial media outlets and abnormal stock returns. Taken together, these findings are consistent with the view that the English language media, and in particular the financial media, are 'propagators of legitimacy' of corporate governance news (see Pollock and Rindova 2003; Dyck et al. 2008; Bushee at al. 2010; Deephouse and Heugens 2009). From a limited attention stance, our findings are consistent with the notion that language barriers inhibit the transmission of news, preventing investors from making fully informed trading decisions (see Ferguson 2015; Engelberg and Parsons 2011).

5.5 Conclusion

In this study we examine the relation between foreign language, media coverage of news and investors' trading behavior. Our first finding suggests that the English language media rely upon foreign-language (Chinese) media outlets to determine the salience of foreign corporate governance news. We attribute this finding to the high search and translation costs associated with processing foreign language news. Our second finding relates to investors' trading behavior post the publication of corporate governance news. While we find that investors react to corporate governance news published by English media outlets, and in particular financial media, corporate governance news reported by foreign (Chinese) media outlets appears to be largely overlooked by investors.

Our findings have important practical implications for the reputational bonding hypothesis. Our evidence suggests that foreign language corporate governance news may be overlooked by U.S. investors simply because the information is not salient. One implication of our findings is that regulators and investors may benefit from employing automated techniques to retrieve and translate news rather than relying upon the English language media to determine the 'newsworthiness' of information.

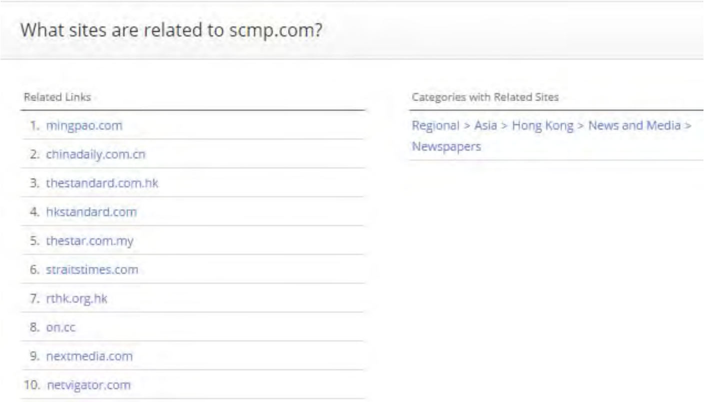
Appendix I

In this section we describe the algorithm to classify media sources into four distinct outlets: *Financial media* comprise of investment focused outlets, *Newswires* comprise of press releases and wire services, *Newspapers* includes daily national and local newspaper outlets and *Other* comprises of blogs and unclassified media sources.

In a first step, an automated online query searches through pages in the web analytics portal alexa.com to retrieve metadata for each media source. Figure 5.1 displays the metadata (in this case a Newspapers tag) for one of the media sources.

Figure 5.1: Illustrative example of a descriptive summary published in Alexa.com

This figure displays the results of a search from the web analytics portal Alexa.com. Metadata for media sources are displayed on the right hand side.



What sites are related to scmp.com?

Related Links

1. mingpao.com
2. chinadaily.com.cn
3. thestandard.com.hk
4. hkstandard.com
5. thestar.com.my
6. straitstimes.com
7. rthk.org.hk
8. on.cc
9. nextmedia.com
10. netvigator.com

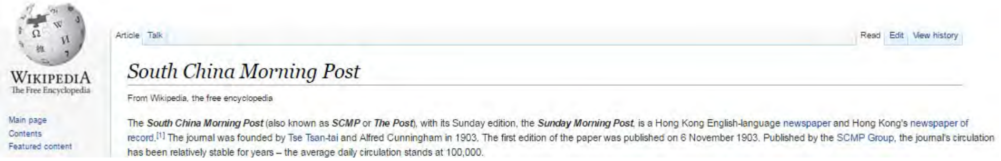
Categories with Related Sites

Regional > Asia > Hong Kong > News and Media > Newspapers

When metadata are not available for a particular media source, the algorithm searches pages in Wikipedia to retrieve descriptive summary text. Figure 5.2 displays an example of the text retrieved for one media source.

Figure 5.2: Illustrative example of a descriptive summary published in Wikipedia

This figure displays the results of a Wikipedia search for one media source. The retrieved text is used to classify the media source into one of four media outlets. Financial media comprise of investment focused outlets, Newswires comprise of press releases and wire services, Newspapers includes daily national and local newspaper outlets and Other comprises of blogs and unclassified media sources.



South China Morning Post

From Wikipedia, the free encyclopedia

The ***South China Morning Post*** (also known as ***SCMP*** or ***The Post***), with its Sunday edition, the ***Sunday Morning Post***, is a Hong Kong English-language newspaper and Hong Kong's newspaper of record.[1] The journal was founded by Tse Tsan-tai and Alfred Cunningham in 1903. The first edition of the paper was published on 6 November 1903. Published by the SCMP Group, the journal's circulation has been relatively stable for years – the average daily circulation stands at 100,000.

In a second step, a multinomial Naïve Bayesian model detects keywords in the Wikipedia text to classify media sources into one of four outlets. Mentions of ‘financ’ and ‘invest’ (in lemmatized form) are used to classify financial media outlets. Text with mentions of ‘newspaper’/‘paper’ and ‘newswire’ are classified as newspaper and newswire outlets respectively.

More formally:

We denote x_{it}, as the number of times word w_t occurs in document D_i. Let $n_i= \Sigma_t x_{it}$ be the total number of words in document D_i. Let $P(w_t|C)$ be the probability that word w_t occurs in class C. We make the simplifying assumption that the probability of each word in a document is independent of the word’s context and position in the document (“bag of words” representation). We can then write the document likelihood $P(D_i|C)$ as a multinomial distribution where the number of draws corresponds to the length of the document, and the proportion of drawing item t is the probability of word type t occurring in a document of class C, $P(w_t|C)$:

$$P(D_i|C) \sim P(x_i|C) = \frac{n_i!}{\prod_{t=1}^{|V|} x_{it}!} \prod_{t=1}^{|V|} P(x_i|C)^{x_{it}}$$

$$\propto \prod_{t=1}^{|V|} P(x_i|C)^{x_{it}} \qquad \textbf{(5.9)}$$

To classify an unlabelled document D_j, we estimate the posterior probability for each class:

$$P(C|D_i) = P(C|x_j)$$

$$\propto P(x_j|C)P(C)$$

$$\propto P(C) \prod_{t=1}^{|V|} P(w_t|C)^{x_{it}} \qquad \textbf{(5.10)}$$

Chapter 6

Inferring employees' social media perceptions of corporate culture[1]

ABSTRACT To date, investors' efforts to 'look inside' a company have been hampered by a lack of data. Traditional survey-based measures are manual and time-consuming to produce, limited in scope with regards to the number of questions they can ask, the number of companies they can cover and their timeliness to collect and process responses. This study seeks to overcome these limitations by inferring employees' perceptions expressed in social media. We find evidence of a statistically significant relation between employees' perceptions of performance-orientated firms and firms' future earnings surprises. Our study highlights the merits of textual analysis for automated corporate culture analysis and builds on the growing body of evidence which suggests that intangible information is not fully exploited by investors.

[1] This is an extended version of the paper: Moniz, A. and de Jong, F, (2014), Sentiment Analysis and the Impact of Employee Satisfaction on Firm Earnings, Advances in Information Retrieval - 36th European Conference on Information Retrieval Research, ECIR 2014. Springer 2014 Lecture Notes in Computer Science.

6.1 Introduction

"Culture is not just one aspect of the game, it is the game"

- *Lou Gerstner, former CEO of IBM*

In recent years, a growing number of knowledge-based firms have abolished their annual employee performance review processes in the belief that they are time-consuming approaches to ascertain backward looking information. Examples of firms include Accenture, Dell and Microsoft. Instead, these firms have opted for more timely goal-setting processes, providing regular employee feedback with the aim of motivating employees and enhancing the firm's operating performance[2]. Indeed, organizational studies suggest that the extent to which employees are motivated by tasks directly impacts their job satisfaction, self-esteem and sense of contributing towards the organization (Locke 1966; Locke and Latham 1990; Yukl and Latham 1978). In particular, individuals exert more effort and work more persistently to attain difficult goals than they do when they attempt to attain less difficult goals or simply 'do their best'. Despite widespread organizational theories documenting the benefits of corporate culture analysis, investors still have little way to 'look inside' a firm. Consequently, investors face a problem of asymmetric information when seeking to infer the value-relevance of firms' intangible assets (see Guiso et al. 2013). In this study we infer a measure of corporate culture from a publicly available social media dataset. The term 'social media' describes a variety of *"new and emerging sources of online information that are created, initiated, circulated and used by consumers intent on educating each other about products, brands, services, personalities and issues"* (Blackshaw and Nazzaro 2006). Text messages, tweets, Facebook sites, blogs and digital videos enable individuals to create, share and recommend information (Gaines-Ross 2010; Elahi and Monachesi 2012). We retrieve 417,645 blog posts for 2,237 U.S. companies from the career community website Glassdoor.com. We draw upon a Natural Language Processing (NLP) technique to infer the latent dimensions of corporate culture. In particular, we infer one dimension which appears to capture employees' perceptions of performance-orientated corporate cultures. We examine the relation between this inferred 'topic cluster' and firms' subsequent earnings surprises. Our findings are consistent with the "errors-in-expectations" hypothesis suggesting that financial analysts systematically underestimate the tangible benefits of corporate culture (cf. Edmans 2011).

We provide three important contributions to the literature. First, we provide a methodology to infer corporate culture from social media. Culture is often defined as *"a set of values, beliefs, and norms of behavior shared by members of a firm that influences individual employee preferences and*

[2] http://fortune.com/2015/10/29/microsoft-dell-performance-reviews/

behaviors" (Besanko et al. 2000). The intangible nature of corporate culture has generated much controversy regarding the creation of a valid construct (Cooper et al. 2001; Pinder 1998; Ambrose and Kulik 1999; O'Reilly et al. 1991). Prior organizational literature either relies upon measures that lack sufficient depth or contain substantial measurement errors (Waddock and Graves 1997; Daines et al. 2010). In recent years, the development of NLP techniques has enabled researchers to automatically organize, summarize, condense unstructured text data, and extract key themes from vast amounts of data. Our approach provides a means to infer employee sentiment at a higher frequency and for a broader cross-section of companies than possible using survey-based measures.

Second, we contribute to literature on investors' underreaction to intangible information. A growing body of research finds that the stock market fails to fully incorporate information regarding a firm's intangible assets (e.g. Edmans 2011; Lev and Sougiannis1996; Chan et al. 2001; Derwall et al. 2011). Under a mispricing channel, an intangible asset only affects the stock price when it subsequently manifests in tangible outcomes which are valued by the stock market. This finding is attributed to the "lack-of-information" hypothesis (Edmans 2011). Our approach is intended to encourage further research into the measurement of intangible information.

The remainder of this study is organized as follows. Section 6.2 outlines related literature, drawing upon goal-setting theory to develop the link between employee motivation and a firm's operating performance. Section 6.3 describes the social media corpus. Section 6.4 provides an overview of probabilistic topic modeling and computational techniques employed to infer corporate culture. Section 6.5 assesses the relation between corporate culture and firms' earnings. Section 6.6 concludes.

6.2 Literature review

6.2.1 Goal setting theory

A number of studies have shown that employees are motivated by specific and challenging objectives and goals (Spector 2003; Bellenger et al. 1984; Coster 1992). Goals motivate high performance by focusing employees' attention, increasing effort and persistence, and encourage innovative solutions to address difficult tasks (Locke and Latham 1990). Goal setting theory suggests that specific rather than abstract goals increase performance. Difficult goals, when accepted by employees, result in higher firm productivity (Latham and Locke 1984). From an employee's perspective, challenging goals often lead to valuable rewards such as recognition, promotions, and/or increases in income from one's work (Latham and Locke 2006). Attaining goals creates a heightened sense of efficacy (personal effectiveness), self-satisfaction, positive affect, and sense of well-being (Wiese and Freund 2005), which in turn increases employee commitment (Tziner and Latham 1989), and reduces staff turnover (Wagner 2007).

6.2.2 Value relevance of intangible information

Traditionally, investors' abilities to decipher the "value relevance" of a firm's intangible assets (such as corporate culture) have been hampered by the lack of data availability. For instance, a firm's human capital management policies must be inferred from Corporate Social Responsibility (CSR) reports (Kolk 2008) or from external surveys such as Fortune's "100 Best Companies to Work for in America" list (Edmans 2011; Levering et al. 1984). These sources suffer from a number of drawbacks. First, CSR disclosures are voluntary in nature and firms' motivations for publishing such disclosures are often unclear. Recent evidence suggests that firms publish CSR reports merely for symbolic purposes to bolster their social images with consumers (Marquis and Toffel 2012; McDonnell and King 2013) rather than to increase transparency and accountability to investors (Moniz and de Jong 2015). In the case of Fortune's Best Places to Work For List, firms pay to participate in the survey which creates perverse incentives for firms to manipulate survey responses (Popadak 2013). Second, CSR may be endogenous with respect to a firm's financial performance - companies may only publish CSR reports if they are more profitable or expect their future profitability to be higher. This relation may hinder investors' abilities to disaggregate the value-relevance of non-financial information (Flammer 2013b). Third, CSR disclosures may be subject to a selection bias if firms' discussions of CSR topics are influenced by institutional pressures (Marquis and Toffel 2012). For instance, prior studies document that non-governmental organizations (NGOs) often choose to scrutinize Wal-Mart's labor relations policies (Bhatnagar 2004; Lobel 2007; Tilly 2007; Rao et al. 2011), and Nike's working conditions (Locke et al. 2007; Greenberg and Knight 2004). Thus NGOs' lobbying pressures may bias the topics discussed in disclosures and hinder investors' abilities to make direct comparisons across firms (Marquis and Toffel 2012). Fourth, CSR reports are often seen as a 'relatively low priority for companies' (Gray et al. 1995a). Firms typically publish CSR disclosures with substantial delay versus accounting-related information, limiting the relevance of the disclosures for investment decisions (Kolk 2008). While survey-based measures of corporate culture seek to address some of these limitations they too suffer from a number of drawbacks. Surveys are typically manually constructed and thus limited in scope by the number of questions they can ask, the number of companies they can cover and suffer in their timeliness to collect and process responses. In the case of Fortune's survey, the results are published infrequently (annually), limited to 100 firms of which only around half are publicly-listed (Popadak 2013), and only composite scores are published potentially obscuring useful information within the overall construct (cf. Daines et al. 2010).

The textual analysis of social media datasets seeks to overcome many of these drawbacks and offers a significant advancement for corporate culture analysis (Popadak 2013). Nonetheless, text poses its own set of analytical challenges. The high costs associated with gathering, processing and structuring text into a standardized format for analysis suggests that intangible information may be overlooked by investors compared to more structured datasets. Thus, even if intangible information is publicly available, it may be ignored by investors if it is not salient (Edmans 2011).

6.3 Data and sample construction

In this section we describe our social media corpus and discuss the challenges associated with automated cultural analysis.

6.3.1 Description of online career community websites

We retrieve employee reviews posted to Glassdoor.com. While there are a number of different career community websites (see Popadak 2013 for a review), Glassdoor.com appears to attract the most diverse set of users. For example, one alternative website provider identifies that its average user is 43 years old with an annual income of $106,000. A second website indicates that its niche market is college students and young professionals (Popadak 2013). By contrast, Glassdoor.com has an estimated 19 million unique users each month and appears to benefit from the most diverse audience as suggested by web traffic statistics from Quantcast.com. The website specializes in audience measurement and employs tracking software to build a picture of web audiences[3]. Table 6.1 reports descriptive statistics on the average profile of Glassdoor users. Profiles appear to be fairly distributed across different sections of society in terms of age, income, education and ethnicity, suggesting that Glassdoor reviews are likely to provide a representative cross-section of an average employee's perceptions within a firm.

[2] http://www.theguardian.com/technology/2012/apr/23/quantcast-tracking-trackers-cookies-web-monitoring

Table 6.1: Descriptive statistics of Glassdoor.com user profiles

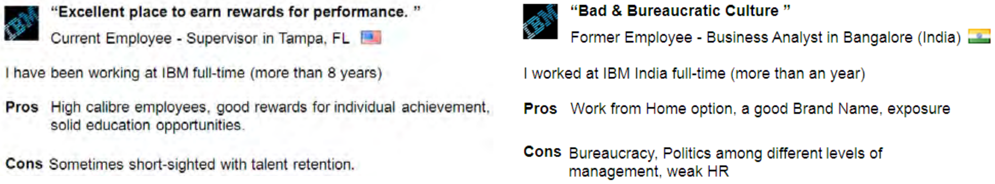
This table reports descriptive statistics of Glassdoor.com user profiles obtained from the web analytics portal quantcast.com as at June 2015. The website measures audience data and compiles visitor profiles by installing tracking pixels on the pages of websites. User profiles include data on gender, age, household income, education level and ethnicity.

Characteristic	Category	Percentage of web traffic
Gender	Male	50%
	Female	50%
Age	< 18	11%
	18-24	18%
	25-34	25%
	35-44	20%
	45-54	17%
	55-64	7%
Household Income	65+	2%
	$0-50k	47%
	$50-100k	30%
	$100-150k	13%
	$150k+	10%
Education Level	No College	27%
	College	51%
	Grad School	22%
Ethnicity	Caucasian	65%
	African American	13%
	Asian	10%
	Hispanic	10%
	Other	2%

Glassdoor states that its website editors seek to ensure the publication of honest, authentic and balanced reviews. Each review must meet strict community guidelines before it is published. Reviewers are required to provide commentary on both the 'pros' and 'cons' of a company to ensure a balanced profile (illustrated in Figure 6.1). Comments are reviewed by website editors to prevent reviewers from posting defamatory attacks, repeat comments or fake reviews while identities are anonymized to allay employees' concerns of company reprisals in the case of negative comments (Popadak 2013). Approximately 15% of reviews are rejected by the website editors because they do not meet their guidelines.

Figure 6.1: Illustrative examples of Glassdoor reviews

This figure illustrates two examples of employee reviews written for IBM. Each review contains metadata which identifies whether a reviewer is a current or former employee, the employee's job title, location and number of years' service at the company. Each review must meet strict community guidelines before it is published. Reviewers are required to provide commentary on both the 'pros' and 'cons' of a company to ensure a balanced profile.



"Excellent place to earn rewards for performance. "

Current Employee - Supervisor in Tampa, FL

I have been working at IBM full-time (more than 8 years)

Pros High calibre employees, good rewards for individual achievement, solid education opportunities.

Cons Sometimes short-sighted with talent retention.

"Bad & Bureaucratic Culture "

Former Employee - Business Analyst in Bangalore (India)

I worked at IBM India full-time (more than an year)

Pros Work from Home option, a good Brand Name, exposure

Cons Bureaucracy, Politics among different levels of management, weak HR

A further advantage of the corpus is a rich set of metadata. The corpus includes the publication date stamp of each review, the number of years' work experience, job title, employment status (part-time/full-time), and work location. In addition, reviewers summarise their opinions of firms in the form of 'star ratings' (on a scale of 1-5). Firms are rated along six dimensions - Culture & Values, Work/Life Balance, Senior Management, Comp & Benefits, Career Opportunities and an Overall Score. In this study we rely upon reviewers' text-based comments which are available from 2008 onwards. By contrast, the star ratings are only available on a consistent basis from 2012. We separately use the star ratings to develop a language-independent measure of sentiment to validate the integrity of our corpus (discussed in Section 6.3.3).

6.3.2 Matching firms to articles

One of the primary challenges associated with unstructured data retrieval is the need to match reviewers' texts to structured data stored in traditional financial databases. We design an algorithm which matches company names in Glassdoor reviews to the CRSP database. The approach takes into account the 'synonym detection problem' typically encountered when matching company names in text (see Engelberg 2008). For instance, the official company name International Business Machines in the CRSP database is more commonly referred to as IBM in Glassdoor reviews. Our algorithm first detects companies' popular names from companies' websites and Wikipedia, then uses this list of names to trawl through Glassdoor.com's subdomains to retrieve relevant reviews. In total, we retrieve 417,645 reviews for 2,237 U.S. companies for the period 2008-2015. Table 6.2 displays descriptive statistics for the sample dataset. Panel A shows that the number of reviews has steadily increased over time. The majority of reviews are posted by North American employees. This observation mitigates a potential concern for cross-cultural analysis that differences in regional locations may account for differences in employees' perceptions (see Hofstede 1980; Triandis et al. 1988). Panel B shows that

60% of the sample consists of reviews posted by individuals stating that they are current rather than former employees. Only a minority (6.9%) of reviewers state that they are part-time employees. Panel C indicates that reviewers have worked in their companies for an average of 1-3 years. Finally, Panel D reports the coverage of reviews by sector using GICS classifications retrieved from the Compustat database. While the corpus includes reviews from all sectors, just over half of the reviews are from the Information Technology and Consumer Discretionary sectors. We view this as a potential benefit of our corpus as these service-based sectors are typically associated with knowledge-based assets such as R&D, human and organizational capital (see Lev 2001).

Table 6.2: Summary statistics for the Glassdoor.com dataset

Panel A: Overview of the dataset by employment location

This table provides descriptive statistics for Glassdoor reviews by employee location (region) and the year the review was posted. This information is obtained from reviwers' metadata. Regions are standardized using MSCI classifications available from https://www.msci.com/market-classification

Region	2008	2009	2010	2011	2012	2013	2014	2015	Total	% of Total
Asia	189	285	926	1,330	6,311	6,264	7,798	2,551	25,654	6%
Europe	307	257	796	435	1,196	1,849	2,949	1,619	9,408	2%
North America	13,139	10,136	15,637	18,068	30,100	45,821	71,444	25,698	230,043	55%
Other	40	53	97	130	632	751	967	429	3,099	1%
Anonymous	1,537	5,001	11,760	13,798	20,931	25,552	46,429	24,433	149,441	36%
Total	15,212	15,732	29,216	33,761	59,170	80,237	129,587	54,730	417,645	
% of Total	3.6%	3.8%	7.0%	8.1%	14.2%	19.2%	31.0%	13.1%	100.0%	100.0%

Panel B: Overview of the dataset by employment status

This table provides descriptive statistics for Glassdoor reviews by employee location and employment status (full-time/part-time, current/former employee). The Anonymous category refers to reviews where employment status was not provided.

Region	Full-time employee	Part-time employee	Anonymous	Total	Current employee	Former employee	Total reviews
Asia ex Japan	18,954	224	6,276	25,454	17,228	8,226	25,454
EMEA	1,121	49	388	1,558	956	602	1,558
Europe	5,787	296	3,325	9,408	6,038	3,370	9,408
Japan	118	9	73	200	109	91	200
Latin America	1,124	18	399	1,541	987	554	1,541
North America	119,506	21,290	89,247	230,043	133,114	96,929	230,043
Anonymous	56,464	6,798	86,179	149,441	93,480	55,961	149,441
Total	203,074	28,684	185,887	417,645	251,912	165,733	417,645
% of Total	48.6%	6.9%	44.5%	100.0%	60.3%	39.7%	100.0%

Panel C: Overview of the dataset by employees' years of service

This table provides descriptive statistics for Glassdoor reviews by location and employees' number of years of service. The number of years of service for an employee is obtained by conducting a textual analysis of reviewers' metadata. The Anonymous category refers to posts where the number of years of service is not provided.

Region	<1 year experience	1-3 years' experience	5+ years' experience	10+ years' experience	Anonymous	Total	% of Total
Asia	3,675	12,800	3,591	591	4,997	25,654	6%
Europe	1,254	3,422	1,348	628	2,756	9,408	2%
North America	33,242	69,275	30,072	17,326	80,128	230,043	55%
Other	330	1,384	616	185	584	3,099	1%
Anonymous	7,829	24,252	13,282	7,383	96,695	149,441	36%
Total	46,330	111,133	48,909	26,113	185,160	417,645	
% of Total	11.1%	26.6%	11.7%	6.3%	44.3%	100.0%	100.0%

Panel D: Overview of the dataset by employment sector

This table provides descriptive statistics for Glassdoor reviews by employment sector and employment status (full-time/part-time, current/former employee). Sectors are retrieved using GICS classifications obtained from the Compustat database.

Sector	Current employee	Former employee	Total reviews	% of Total	Number of unique firms
Energy	5,753	3,445	9,198	2.2%	113
Materials	4,295	2,670	6,965	1.7%	110
Industrials	27,616	18,150	45,766	11.0%	317
Consumer Discretionary	54,387	45,415	99,802	23.9%	343
Consumer Staples	14,736	9,560	24,296	5.8%	92
Health Care	16,643	11,488	28,131	6.7%	309
Financials	25,600	18,318	43,918	10.5%	317
Information Technology	89,197	46,903	136,100	32.6%	482
Telecommunication Services	2,359	1,521	3,880	0.9%	28
Utilities	1,734	936	2,670	0.6%	52
Unclassified	9,592	7,327	16,919	4.1%	74
Total	251,912	165,733	417,645	100.0%	2,237

Taken together, we believe that our corpus offers a more timely way to conduct corporate culture analysis across a large cross-section of companies compared to more traditional survey-based measures.

6.3.3 Validating the corpus

One of the challenges associated with the retrieval of online reviews is the potential for sample bias. The bias refers to the potential to select a misrepresentative sample of reviews which may hinder statistical inference. In particular, differences in reviewers' native languages, cultures and human emotional experiences may result in unintended consequences when inferring sentiment (see Hogenboom et al. 2012; Pang and Lee 2004; Wierzbicka 1995). For instance, disgruntled former employees may have greater incentive to post negative comments about their previous employers.

To assess the potential for sampling bias, we draw upon NLP literature and employ an approach which seeks to compare the information expressed in reviewers' star ratings versus their texts (Hogenboom et al. 2012, 2014). A "language-independent" measure of sentiment starts from the premise that star ratings are universal classifications of a reviewer's intended sentiment, independent of potential language, cultural or emotional differences. Regardless of a reviewer's background, we expect to observe a monotonically increasing relationship between a reviewer's star rating and the expression of sentiment as inferred from text. We estimate a panel regression where the dependent variable is the Overall star rating for a firm and the independent variables are features extracted from reviewers' texts. Company fundamental data are retrieved from standard financial databases. Price related variables are obtained from CRSP; accounting data are obtained from COMPUSTAT and analyst information from I/B/E/S. For controls, we include the star ratings: COMP is the ‘Comp & Benefits’ star rating, WORKLIFE is the ‘Work/Life Balance’ rating, MGT is reviewers’ ‘Senior Management’ rating, CULTURE is the ‘Culture & Values’ star rating and CAREER is the ‘Career Opportunities’ rating. We supplement the star ratings with two indicator variables. The variable, *Part-time*, equals one if a reviewer is a part-time worker. The variable, *Former*, equals one if a reviewer is a former employee. These features are identified using reviewers’ metadata. To control for a potential ‘halo’ effect caused when reviewers implicitly form their perceptions using publicly available information (see Fryxell and Wang 1994; Brown and Perry 1994), we include controls for book-to-market (Log(Book/Market)), analyst revisions (Analyst Revisions), price momentum (Pmom) and one-year historic sales growth (SG). Log(Book/Market)) is the natural log of the book-to-value of equity measured as at the end of the preceding calendar year, following Fama and French (1992). Analyst revisions is the 3-month sum of changes in the median analyst’s forecast, divided by the firm’s stock price in the prior month (Chan et al. 1996). Pmom is the (signed) stock’s return measured over the previous 12 months. Finally, we include firm size (Log(Market Equity)) measured at the end of the preceding calendar year. Table 6.3 displays the regression results.

Table 6.3: Regression of reviewers' Overall star ratings

This table reports regression results on the relation between reviewers' Overall ratings, employees' metadata and firm characteristics. The dependent variable is Overall star rating score provided by Glassdoor reviewers (scale 1-5). Former is an indicator variable equal to one if the reviewer is a former employee of the company, and zero otherwise. Part-time is an indicator variable equal to one if the reviewer is a part-time worker, and zero otherwise. Please refer to the text for a description of the control variables. For presentational reasons, the star ratings 'Comp & Benefits', 'Work/Life Balance', 'Senior Management', 'Culture & Values' and 'Career Opportunities', included as control variables are hidden from the table. Standard errors are clustered by firm following Petersen (2009). For each variable we report the corresponding robust t-statistic (in parentheses). Sample period: 2008-2015.

	(1)	(2)
Intercept	0.006	0.040
	(2.320)	(1.657)
Former	-0.058	-0.059
	(-4.863)	(-4.754)
Part-time	0.053	0.053
	(5.296)	(5.875)
Log(Book/Market)		0.004
		(1.663)
Log(Market Equity)		0.000
		(2.099)
Analyst revisions		0.632
		(2.312)
SG		0.006
		(0.170)
Pmom		0.009
		(1.417)
R^2	0.734	0.744

The regressions indicate that reviewers' star ratings are, on average, significantly lower for former employees and significantly higher than average for part-time employees. To limit the potential for sample bias, we choose to exclude these two groups of reviews from the corpus. Although this approach reduces the number of observations in our dataset, it allows for more meaningful recommendations for corporate culture analysis based upon an analysis of firms' current, full-time employees.

6.4 Inferring corporate culture

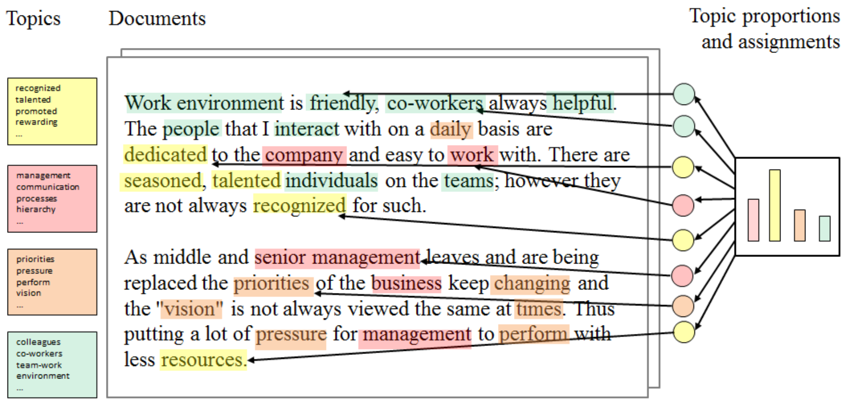
In this section we describe the probabilistic topic model used to infer corporate culture. We then examine the characteristics of performance-orientated firms.

6.4.1 Topic model of corporate culture

Topic models are statistical models that posit low-dimensional representations of data to infer latent topics in text. The most common topic modeling approach is Latent Dirichlet Allocation (LDA) (Blei et al. 2003). The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The model is based on the hypothesis that an employee writing a document has certain topics in mind. To write about a topic means to pick a word with a certain probability from the pool of words of that topic. A whole document can then be represented as a mixture of different topics. In the case of the Glassdoor corpus, these topics reflect the employee's view of a company and his/her particular vocabulary. Topics can be thought of as "semantically related probabilistic clusters of words". For instance, a discussion about an employee's working environment may include references to 'colleagues', 'co-workers', and 'teams'. By contrast, a discussion about employee performance may include the terms: 'recognition' and 'promotion'. The goal of topic modeling is to automatically discover these topics from a collection of documents. While the documents are observed, the topic structure (the topics, per-document topic distributions, and the per-document per-word topic assignments) are hidden structure (Blei et al. 2003). Figure 6.2 provides an extract of an employee review to illustrate the methodology.

Figure 6.2 Illustrative example of LDA for an employee review

This figure has been adapted from Blei (2012) and is intended to illustrate the premise of probabilistic topic modelling. LDA assumes that a number of topics which are distributions over words exist for the whole collection (far left). Each document is assumed to be generated as follows: First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored circles) and choose the word from the corresponding topic.


Topics
Documents
Topic proportions and assignments
recognized talented promoted rewarding ...
management communication processes hierarchy ...
priorities pressure perform vision ...
colleagues co-workers team-work environment ...
Work environment is friendly, co-workers always helpful. The people that I interact with on a daily basis are dedicated to the company and easy to work with. There are seasoned, talented individuals on the teams; however they are not always recognized for such.
As middle and senior management leaves and are being replaced the priorities of the business keep changing and the "vision" is not always viewed the same at times. Thus putting a lot of pressure for management to perform with less resources.

We employ a variant of LDA to detect a specific one dimension of corporate culture, namely employees' perceptions of performance-orientated firms (see Appendix I for a description of the methodology). The output of the algorithm is a percentage which reflects the degree to which a given reviewer discusses performance-orientated aspects in text. We label this measure GOAL.

To distinguish between positive and negative sentiment reviews we employ a variant of the 'term counting' measures of sentiment commonly employed by the extant financial literature (see Tetlock 2008). Specifically, we compute the fraction of the number of positive (P) versus negative (N) terms in each review versus a pre-defined lexicon of terms (the General Inquirer dictionary), and adjust[4] for negation terms in text. Negation terms can alter the polarity of the words or phrases they precede and are often seen in social media texts due to the more frequent use of colloquialisms (see Dadvar et al. 2011; Hu and Liu 2004). We label measure of reviewer sentiment, TONE.

[4] The approach ignores the polarity of a term matched in the General Inquirer dictionary if there is a negation term (such as 'although', 'but', 'no', 'not', 'yet') within five words of it.

Table 6.4: Illustrative examples of employees' perceptions of performance-orientated cultures

This figure provides illustrative examples of employee reviews associated with performance-orientated cultures. To aid readers' interpretation, we randomly select five comments with positive and negative sentiment. Sentiment is computed by counting the number of positive (P) versus negative (N) terms, adjusting for negation terms, using the General Inquirer dictionary (Stone et al. 1966). The spelling mistakes and grammatical errors are as published in the online reviews.

Panel A: Examples of positive sentiment reviews

Good foundation in place, with a common goal to understand everyone
if your hard working, it's a good place to work. It weeds out the lazy people and the people that dont want to work.
Good people that have same goal.
Great place to work if you are not lazy.
Well planned work habits, good company culture.

Panel B: Examples of negative sentiment reviews

The most hardest thing here is hitting your numbers. If you don't reach the desired goal of the company, they will get rid of you.
Not a very good work life balance and aggressive deadlines.
The fact that the end goal of JPM is always bottom line, the workload and hours are very intense but the work is exciting and worth it.
Fast pace and high stress of goal for achievement and sucess.
Long work hours, stressful sometimes, had to work in weekends to meet deadlines.

A manual inspection of reviews appears to illustrate that the topic model detects both direct outcomes associated with performance-orientated cultures, such as references to 'goals', as well as indirect outcomes such as 'pressure' and 'stress' (see Table 6.4).

6.4.2 Data and summary statistics

To align the frequency of accounting data with reviewers' texts, we aggregate comments between firms' successive earnings announcement dates and create a composite document per firm per quarter. We winsorize firm characteristics at the 1% level to eliminate the impact of outliers. Panel A of Table 6.5 reports the median fundamental characteristics of firms when sorted into quartiles by their GOAL score. The last column illustrates the statistical significance of a difference of means t-test between top and bottom quartile firms for each fundamental characteristic. Companies with reviews in the highest GOAL quartile exhibit significantly higher growth than firms in the lowest quartile. This finding is consistent across asset, employee, and sales growth. Panel B of Table 6.5 reports the Spearman rank correlations between the average Glassdoor star ratings and GOAL. The correlations between GOAL and the overall composite star rating appear to be relatively low suggesting that the perceptions inferred from GOAL differ from the information provided by reviewers' star ratings.

Table 6.5: Descriptive statistics of firm characteristics

Panel A: Fundamental characteristics

This table reports the median fundamental characteristics of firms when sorted into quartiles by their GOAL score. GOAL is the proportion of reviews that refer to performance-orientated cultures as inferred by the LDA topic model. We create a composite document per firm per quarter to align the reporting frequency of accounting variables (quarterly/annual) with reviewers' comments (daily) by aggregating reviewer comments between earnings announcement dates (sourced from IBES). We winsorize all firm characteristics at the 1% level to eliminate the impact of outliers. OVERALL is the overall star rating score provided by Glassdoor reviewers, averaged between consecutive earnings announcement dates per company. All fundamental data comes from COMPUSTAT Fundamentals Annual Database. The final column of the table indicates the statistical significance of a difference of means t-test between top and bottom quartile firms for each fundamental characteristic where *** indicates statistical significance at the 1% level, ** at the 5% level and * at the 10% level. Sample period: 2008-2015.

Characteristic	**1st.Quartile**	**2nd.Quartile**	**3rd.Quartile**	**4th.Quartile**	**Diff of means T-test (Q1 vs. Q4)**
Accruals	-0.044	-0.035	-0.043	-0.042	
Asset growth (yoy)	0.037	0.051	0.086	0.087	***
Employee growth (yoy)	0.021	0.028	0.046	0.052	***
Financial leverage	0.429	0.510	0.318	0.307	
Market capitalisation (US$ mn)	13,329	17,178	22,289	28,408	***
Prior price momentum	0.148	0.164	0.150	0.198	*
ROA	0.146	0.149	0.149	0.162	***
Sales growth (yoy)	0.038	0.045	0.057	0.069	***
Tobin's Q	1.329	1.474	1.620	1.837	***
GOAL	0.040	0.072	0.099	0.145	
OVERALL	3.231	3.308	3.505	3.500	***

Panel B: Correlation of Glassdoor.com star rating scores

This table reports the Spearman rank correlations between the star ratings provided by Glassdoor reviews and GOAL. GOAL is the proportion of reviews that refer to performance-orientated cultures as inferred by the LDA topic model. OVERALL is the overall star rating score provided by Glassdoor reviewers. COMP is the 'Comp & Benefits' star rating provided by Glassdoor reviewers. WORKLIFE is the Glassdoor 'Work/Life Balance' rating. MGT is reviewers' 'Senior Management' rating. CULTURE refers to the 'Culture & Values' star rating and CAREER is the Glassdoor 'Career Opportunities' rating. Sample period: 2008-2015.

	GOAL	OVERALL	COMP	WORKLIFE	MGT	CULTURE	CAREER
GOAL	1.00						
OVERALL	0.04	1.00					
COMP	0.12	0.58	1.00				
WORKLIFE	0.05	0.76	0.56	1.00			
MGT	0.01	0.74	0.44	0.63	1.00		
CULTURE	0.00	0.60	0.41	0.49	0.54	1.00	
CAREER	0.03	0.76	0.54	0.74	0.65	0.49	1.00

6.4.3 Validating the GOAL measure

Next we regress firms' GOAL scores on Glassdoor 'star ratings' to examine whether the information inferred from reviewers' texts is incremental to the information provided in star ratings and TONE. We include controls for book-to-market (Log(Book/Market)), analyst revisions (Analyst Revisions), price momentum (Pmom) and one-year historic sales growth (SG). Log(Book/Market)) is the natural log of the book-to-value of equity measured as at the end of the preceding calendar year, following

Fama and French (1992). Analyst revisions is the 3-month sum of changes in the median analyst's forecast, divided by the firm's stock price in the prior month (Chan et al. 1996). Pmom is the (signed) stock's return measured over the previous 12 months, while firm size (Log(Market Equity)) is measured at the end of the preceding calendar year. We control for firms' corporate social responsibility attributes to assess the claim that performance-orientated firms undertake unethical behavior due to the high financial incentives associated with meeting performance targets (Jensen 2002; Schweitzer et al. 2004; Ordóñez et a. 2009). Following prior studies (see Waddock and Graves 1997; Hillman and Keim 2001; Statman and Glushkov 2009), we proxy this behavior by including an "employee relations" metric obtained from the KLD database. In line with standard practice, we calculate net employee strengths by summing all identified strengths and subtracting all identified weaknesses in a given year (see Verwijmeren 2010). Finally, we include an employee satisfaction measure to evaluate whether the characteristics of performance-orientated firms differ from the information published in Fortune magazine's "100 Best Companies to Work for in America" list (Edman 2011). We create an indicator variable, BC, equal to one if a company is listed in the Fortune list, and zero otherwise. Following Petersen (2009), standard errors are clustered by firm to correct for time series dependence in standard errors. Table 6.6 reports the regression results.

Column 2 identifies a positive correlation between GOAL and Glassdoor star ratings for management quality and opportunities, while a negative relation between GOAL and compensation. These findings are consistent with the view that performance-orientated firms seek to incentivize individuals by providing a larger proportion of their total compensation in variable pay (Gneezy et al. 2011; Kamenica 2012), making the fixed component relatively unattractive versus competitors (Gerhart and Rynes 2003; Adams 1963). Column 2 also identifies a positive correlation to one-year historic sales growth and price momentum, indicating that performance-orientated firms are typically growth companies. Finally, Column 3 controls for CSR metrics and suggests that GOAL is not subsumed by employee relations or employee satisfaction. Taken together our findings suggest that GOAL is a distinct dimension of corporate culture.

Table 6.6: Regression of GOAL and firm characteristics

This table reports the relation between GOAL and company characteristics. The dependent variable is the topic probability associated with goal-setting behavior inferred by the LDA model. OVERALL is the Glassdoor Overall star rating provided by Glassdoor reviewers, COMP is the 'Comp & Benefits' star rating. WORKLIFE is the Glassdoor 'Work/Life Balance' rating. MGT is reviewers' 'Senior Management' rating, CULTURE refers to the 'Culture & Values' star rating and CAREER is the Glassdoor 'Career Opportunities' rating. TONE is a measure of document polarity computed by counting the number of positive (P) versus negative (N) terms using the General Inquirer dictionary (Stone et al. 1966). Log (Book/Market) is the natural log of the book-to-value of equity as of the previous year end. SG is one-year sales growth. Analyst revisions is the 3-month sum of changes in the median analyst's forecast, divided by the firm's stock price in the prior month. ROA is net income before depreciation scaled by total assets as at the previous year end. Pmom is the (signed) stock's return measured over the previous 12 months. SG is one-year sales growth. The fundamental data comes from COMPUSTAT Fundamentals Annual Database apart from Analyst revisions which comes from I/B/E/S and Pmom from CRSP. KLD is a measure of employee relations metric obtained from the KLD database and is defined as the difference between employee strengths and concerns over the past year. BC is an indicator variable equal to one if the company is in Fortune magazine's "100 Best Companies to Work for in America" list, and zero otherwise (Edmans 2011). Standard errors are clustered by firm following Petersen (2009). For each variable we report corresponding robust t-statistic (in parentheses). Sample period: 2008-2015.

	(1)	(2)	(3)
OVERALL	0.012	0.078	0.011
	(2.867)	(5.663)	(2.63)
TONE	0.023	0.026	0.009
	(0.798)	(0.986)	(0.312)
COMP		-0.047	
		(-11.505)	
WORKLIFE		-0.002	
		(-1.92)	
MGT		0.032	
		(5.581)	
CULTURE		0.001	
		(0.308)	
OPPORTUNITIES		0.022	
		(3.853)	
Log(Book/Market)	-0.003	0.001	-0.002
	(-1.078)	(0.463)	(-0.786)
ROA	0.027	-0.004	0.028
	(1.117)	(-0.175)	(1.19)
SG	0.056	0.030	0.055
	(4.213)	(2.39)	(3.986)
Analyst revisions	0.147	0.070	0.095
	(1.21)	(0.634)	(0.721)
Pmom	0.009	0.009	0.007
	(2.288)	(2.724)	(1.874)
KLD			-0.003
			(-2.734)
BC			-0.018
			(-0.885)

6.5 Empirical results

This section investigates the relation between performance-orientated cultures, firm value and firms' future earnings.

We compute Tobin's Q as a measure of firm value, defined as the market value of the firm divided by the replacement value of the firm's assets. The market value of assets is measured as the sum of the book value of assets and the market value of common stock outstanding minus the sum of the book value of common stock and balance sheet deferred taxes. Replacement value is represented by the book value of assets (Kaplan and Zingales 1997). We control for sector, region and year effects and run pooled OLS regressions to estimate models of Tobin's Q. We test for the significance of the coefficients using standard errors that are robust to heteroskedasticity clustered by firm (Petersen 2009). The pooled regression results are reported in Table 6.7.

Table 6.7: Regression of GOAL and firm value

This table reports the results of running quarterly regressions of firm value on a set of independent variables. The dependent variable is Tobin's Q, defined as the market value of the firm divided by the replacement value of the firm's assets. We compute the market value of assets as the sum of the book value of assets and the market value of common stock outstanding minus the sum of the book value of common stock and balance sheet deferred taxes. GOAL is the proportion of reviews that refer to performance-orientated cultures as inferred by the LDA topic model. A composite document is computed for each firm by aggregating Glassdoor reviews between consecutive earnings announcement dates for each firm. Earnings announcement dates are sourced for I/B/E/S. A minimum of 30 reviews are required to create a composite document per firm. OVERALL is the Glassdoor Overall star rating averaged across reviews with the composite document. TONE is a measure of document polarity computed by counting the number of positive (P) versus negative (N) terms, adjusted for negation terms, using the General Inquirer dictionary (see text for details). The definitions for the fundamental variables are described in the text and come from COMPUSTAT Fundamentals Annual Database apart from Analyst revisions which comes from I/B/E/S and Pmom from CRSP. KLD is a measure of employee relations metric obtained from the KLD database and is defined as the difference between employee strengths and concerns over the past year. BC is an indicator variable equal to one if the company is in Fortune magazine's "100 Best Companies to Work for in America" list, and zero otherwise (Edmans 2011). Standard errors are clustered by firm following Petersen (2009). For each variable we report corresponding robust t-statistic (in parentheses). Sample period: 2008-2015.

	(1)	(2)	(3)
GOAL	1.624	1.400	1.720
	(2.691)	(2.023)	(2.823)
TONE	-0.374	-0.034	-0.166
	(-0.701)	(-0.046)	(-0.311)
OVERALL	0.328		0.322
	(4.472)		(4.393)
COMP		-0.211	
		(-1.848)	
WORKLIFE		0.143	
		(1.181)	
MGT		0.261	
		(1.679)	
CULTURE		-0.102	
		(-1.007)	
OPPORTUNITIES		0.300	
		(1.738)	
log(Book/Market)	-0.762	-0.744	-0.699
	(-2.634)	(-2.488)	(-2.69)
ROA	4.348	4.057	4.725
	(3.364)	(3.282)	(3.192)
SG	2.846	2.635	2.736
	(2.941)	(2.921)	(2.255)
KLD			-0.073
			(-3.727)
BC			1.130
			(2.978)

Column 1 indicates a positive and highly statistically significant coefficient for GOAL, suggesting that performance-orientated firms tend to be more profitable. Column 2 indicates that there is no evidence of a statistical relation between the underlying star ratings and firm value. Column 3 indicates that GOAL is incremental to the employee satisfaction and employee relations metrics.

Next we hypothesize that if financial analysts overlook intangible information, potentially due to the costs associated with gathering, processing and analysing unstructured data, we would expect that positive benefits of corporate culture are only recognized once they manifest into tangible outcomes post earnings announcements (see Easterwood and Nutt 1999; Edmans 2011). Our main test computes each firm's standardized unexpected earnings (SUE) using a seasonal random walk with trend model for each firm's earnings (Bernard and Thomas 1989):

$$UE_t = E_t - E_{t-4}$$

$$SUE_t = \frac{UE_t - \mu_{UEt}}{\sigma_{UEt}} \tag{6.1}$$

where E_t is the firm's earnings in quarter t, and the trend and volatility of unexpected earnings (UE) are equal to the mean (μ) and standard deviation (σ) of the firm's previous 20 quarters of unexpected earnings data, respectively.

Following Tetlock (2008), we require that each firm has non-missing earnings data for the most recent 10 quarters and assume a zero trend for all firms with fewer than 4 years of earnings data. We use the median analyst forecast from the most recent statistical period in the I/B/E/S summary file prior to the earnings announcement. We winsorize SUE and all analyst forecast variables at the 1% level to reduce the impact of estimation error and extreme outliers respectively. We create a composite document for each firm to align different frequencies of data by aggregating Glassdoor reviews between consecutive earnings announcement dates. We require a minimum of 30 reviews per company between quarterly earnings announcements to avoid drawing statistical inferences using a limited and potentially unrepresentative set of employee comments. For control variables we include firms' lagged earnings, size, book-to-market ratio, analysts' earnings forecast revisions, and analysts' forecast dispersion. We measure firms' lagged earnings using last quarter's SUE. We compute analysts' forecast dispersion (Forecast Dispersion) as the standard deviation of analysts' earnings forecasts in the most recent time period prior to the earnings announcement scaled by earnings volatility (σ). Finally, we compute an indicator variable, *Difficulty*, to test the hypothesis that difficult and challenging goals are associated with higher firm productivity (Latham and Locke 1984). The indicator is equal to one if GOAL x TONE is negative, and zero otherwise. Examples of terms captured by the indicator include references to 'tough' and 'difficult' goals. Table 6.8 reports the regression results. Standard errors are clustered by calendar quarter (following Petersen 2009).

Column 2 identifies a positive and highly statistically significant coefficient for GOAL, suggesting that the measure contains incremental information for predicting earnings surprises beyond those of company fundamentals or TONE. Column 3 suggests that firms which employees perceive to have tough goals are positively associated with future earnings surprises. Column 4 controls for employee relations and satisfaction and suggests that the information contained in GOAL is not subsumed by these measures.

Table 6.8: Regression of GOAL and earnings surprises

This table provides the OLS regression estimates of the relation between GOAL and a firm's one-quarter ahead earnings surprise (SUE). The dependent variable, SUE, is a firm’s standardized unexpected quarterly earnings. GOAL is the proportion of reviews that refer to performance-orientated cultures as inferred by the LDA topic model. A composite document is computed for each firm by aggregating Glassdoor reviews between consecutive earnings announcement dates for each firm. Earnings announcement dates are sourced for I/B/E/S. A minimum of 30 reviews are required to create a composite document per firm. OVERALL is the Glassdoor Overall star rating averaged across reviews with the composite document. TONE is a measure of document polarity computed by counting the number of positive (P) versus negative (N) terms, adjusted for negation terms, using the General Inquirer dictionary (see text for details). Difficulty is an indicator variable equal to one if the interaction term GOAL x TONE is negative, and zero otherwise. The interaction proxies whether employees' perceptions of performance-orientated cultures are negative (examples of negative terms include 'tough', 'difficult', and 'stress'). KLD is a measure of employee relations metric obtained from the KLD database and is defined as the difference between employee strengths and concerns over the past year. BC is an indicator variable equal to one if the company is one of Fortune magazine’s “100 Best Companies to Work for in America”, and zero otherwise (Edmans 2011). Regressions include control variables for lagged firm earnings, firm size, book-to-market, trading volume, past stock returns, and analysts’ quarterly forecast revisions and dispersion (see text for details). Standard errors are clustered by firm following Petersen (2009). For each variable we report corresponding robust t-statistic (in parentheses). Sample period: 2008-2015.

	(1)	(2)	(3)	(4)
lagged dependent	-0.012	-0.015	-0.013	-0.012
	(-0.358)	(-0.423)	(-0.38)	(-0.351)
Forecast dispersion	-2.700	-2.806	-2.283	-2.581
	(-3.196)	(-3.318)	(-2.656)	(-2.916)
OVERALL	0.067	0.053	0.061	0.079
	(0.761)	(0.505)	(0.582)	(0.755)
GOAL		1.770	4.665	4.477
		(2.536)	(3.931)	(3.751)
TONE		0.054	1.652	1.714
		(2.071)	(1.735)	(1.796)
Difficulty			14.780	14.180
			(3.008)	(2.892)
Analyst revisions	15.130	14.730	13.910	18.050
	(4.749)	(4.622)	(4.369)	(5.173)
Log(Market Equity)	0.000	0.000	0.000	0.000
	(-1.078)	(-1.021)	(-1.183)	(-1.552)
Log(Book/Market)	-0.006	-0.018	0.001	-0.053
	(-0.096)	(-0.294)	(0.009)	(-0.857)
Pmom	0.716	0.738	0.742	0.774
	(7.411)	(7.612)	(7.689)	(8.007)
KLD				0.055
				(1.904)
BC				-0.974
				(-1.699)

6.6 Conclusion

To date, investors' efforts to ‘look inside’ a company have been hampered by a lack of data. Traditional survey-based measures are manual and time-consuming to produce, limited in scope with regards to the number of questions they can ask, the number of companies they can cover and their timeliness to collect and process responses. This study seeks to overcome these limitations by inferring employees’ perceptions expressed in social media. We find evidence of a statistically significant relation between performance-orientated firms and firms' future earnings surprises. Our findings are consistent with the notion of “errors-in-expectations” in financial analysts' forecasts, suggesting that performance-orientated cultures are recognized once the tangible benefits of corporate culture materialize post earnings announcements.

6.7 Appendix I

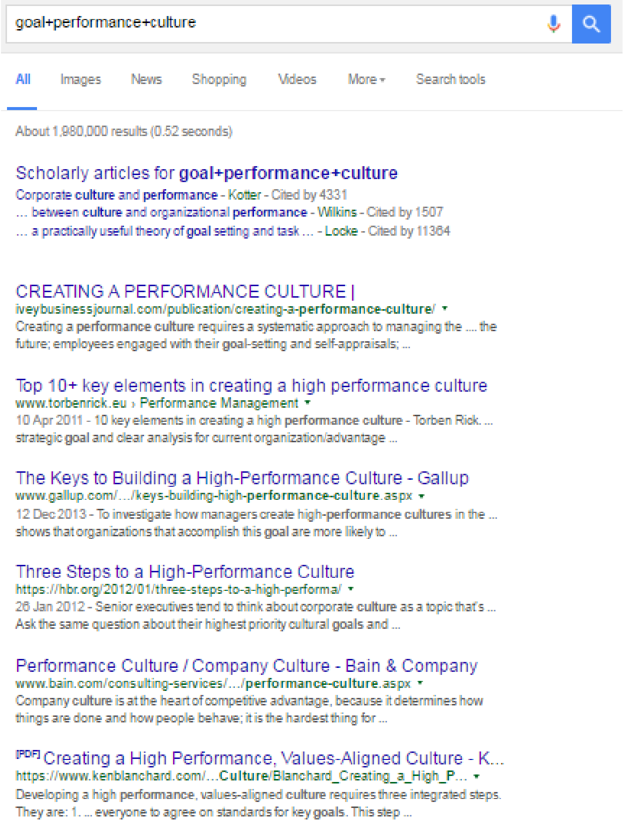
A topic model is a statistical model for learning abstract “topics” in documents. Topic models have played an important role in a variety of data mining tasks, within computer science (Blei et al. 2003; Griffiths and Steyvers 2004; Ramage et al. 2010; Liu et al. 2009), social and political science (Ramage et al. 2009; Grimmer 2010), and humanities (Mimno 2012) for the categorization and summarization of texts. The intuition behind LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA is most easily described by its generative process which models the way documents arise. For each document, we generate words in a two-stage process:

1. Randomly choose a distribution over topics.
2. For each word in the document:
    a. Randomly choose a topic from the distribution over topics in step #1.
    b. Randomly choose a word from the corresponding distribution over the vocabulary.

Each document exhibits topics in different proportions (step #1); each word in each document is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a). Thus, the goal of topic modeling is to automatically discover these latent topics from the collection of documents. The documents themselves are observed, while the topic structure - the topics, per-document topic distributions, and the per-document per-word topic assignments are hidden structure (Blei et al. 2003).

We employ a variant of the standard implementation of LDA and guide the topic model by providing a set of seed words of interest. Prior NLP studies show that seed words can help guide the topic discovery process (see Jagarlamudi et al. 2012). The approach is designed to improve both topic word distributions (by biasing topics to produce appropriate seed words) and to improve document-topic distributions (by biasing documents to select topics). To create a seed word list, we employ an approach which generates terms intended to capture online domain knowledge. In a first step, we enter the query ‘goals+performance+culture’ into an online search engine. While the terms for this search query are chosen heuristically, the approach is designed to detect a greater variety of synonyms to create an ontology of terms associated with performance-orientated cultures. The results of the search query are displayed in Figure 6.3.

Figure 6.3: Display of the search query results for 'goals+performance+culture'
This figure displays the results for the search query 'goals+performance+culture'. The goal is to retrieve a broad selection of online documents to infer domain knowledge regarding corporate culture. The titles and summary texts retrieved from the search results online are used to automatically generate an ontology (sets of words related to the initial search query) a link analysis algorithm.



goal+performance+culture

About 1,980,000 results (0.52 seconds)

Scholarly articles for **goal+performance+culture**
Corporate **culture** and **performance** - Kotter - Cited by 4331
… between **culture** and organizational **performance** - Wilkins - Cited by 1507
… a practically useful theory of **goal** setting and task … - Locke - Cited by 11364

CREATING A PERFORMANCE CULTURE |
iveybusinessjournal.com/publication/creating-a-**performance-culture**/ ▾
Creating a **performance culture** requires a systematic approach to managing the .... the future; employees engaged with their **goal**-setting and self-appraisals; ...

Top 10+ key elements in creating a high performance culture
www.torbenrick.eu › Performance Management ▾
10 Apr 2011 - 10 key elements in creating a high **performance culture** - Torben Rick. ... strategic **goal** and clear analysis for current organization/advantage ...

The Keys to Building a High-Performance Culture - Gallup
www.gallup.com/…/keys-building-high-**performance-culture**.aspx ▾
12 Dec 2013 - To investigate how managers create high-**performance cultures** in the ... shows that organizations that accomplish this **goal** are more likely to ...

Three Steps to a High-Performance Culture
https://hbr.org/2012/01/three-steps-to-a-high-performa/ ▾
26 Jan 2012 - Senior executives tend to think about corporate **culture** as a topic that's ... Ask the same question about their highest priority cultural **goals** and ...

Performance Culture / Company Culture - Bain & Company
www.bain.com/consulting-services/…/**performance-culture**.aspx ▾
Company **culture** is at the heart of competitive advantage, because it determines how things are done and how people behave; it is the hardest thing for ...

[PDF] Creating a High Performance, Values-Aligned Culture - K…
https://www.kenblanchard.com/…**Culture**/Blanchard_Creating_a_High_P… ▾
Developing a high **performance**, values-aligned **culture** requires three integrated steps. They are: 1. ... everyone to agree on standards for key **goals**. This step ...

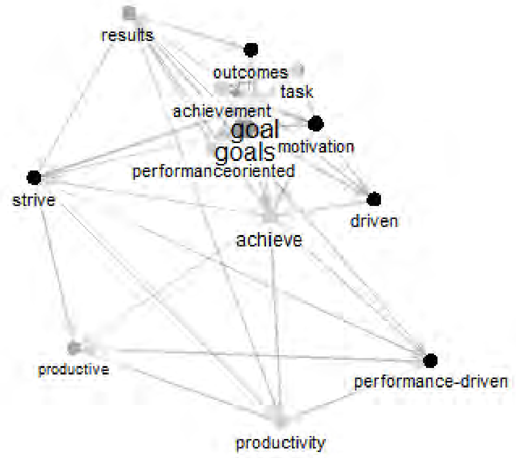
In a second step, we employ a link analysis to detect the most frequent terms mentioned in the titles and summary texts retrieved from the search results. Graph-based algorithms have received much attention (Mihalcea and Tarau 2004) as an approach to keyphrase extraction and are considered to be state-of-the-art unsupervised methods (Liu et al. 2009). In a graph representation of a document, nodes are words or phrases, and edges represent co-occurrence or semantic relations. The underlying assumption is that all words in the text have some relationship to all other words in the text. Such an approach is statistical, because it links all co-occurring terms without considering their meaning or function in text. Centrality is often used to estimate the importance of a word in a document (Opsahl et al. 2010), and is a way of deciding on the importance of a vertex within a graph that takes into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information (Boudin 2013). Specifically, we employ TextRank (Mihalcea and Tarau 2004), a ranking algorithm based on the concept of eigenvector centrality, computes the importance of the nodes in the graph. Each vertex corresponds to a word. A weight, w_{ij}, is assigned to the edge connecting the two vertices, v_i and v_j. The goal is to compute the score of each vertex, which reflects its importance, and use the word types that correspond to the highest scored vertices to form keywords for the text (Boudin 2013). The score for v_i, $S(v_i)$, is initialized with a default value and is computed in an iterative manner until convergence using recursive formula shown in Equation (6.2).

$$S(v_i) = (1 - d) + d\ x \sum_{v_j \in Adj(v_i)} \frac{w_{ji}}{\sum_{v_k \in Adj(v_j)} w_{jk}} S(v_j) \tag{6.2}$$

where $Adj(v_i)$ denotes v_i's neighbors and d is the damping factor set to 0.85 (Mihalcea and Tarau 2004). Figure 6.4 displays the resulting clustering of terms. The size of each node is directly proportional to the TextRank score.

Figure 6.4: Link analysis approach used to generate seed words for the LDA model

This figure displays the results of a link analysis based on detecting the most frequently mentioned terms generated by the online search query: 'goals+performance+culture'. TextRank (Mihalcea and Tarau 2004) a ranking algorithm based on the concept of eigenvector centrality, is employed to compute the importance of the nodes in the graph. The size of each node is directly proportional to the TextRank score. Different nodes colors reflect different communities identified using the Clauset-Newman-Moore algorithm and are for illustrative purposes only.



Chapter 7

Inferring the financial materiality of Corporate Social Responsible news

ABSTRACT We retrieve 105,983 news articles from newswires, newspapers, blogs and magazines over the period 1980-2014 and employ a probabilistic topic model to infer journalists' and other stakeholders' attributions of firms' poor corporate socially responsible (CSR) practices. The approach seeks to automatically detect contextual information and semantic meaning in text. Attributions range from allegations/criticisms over poor CSR to more material concerns reflecting corporate difficulties and litigation risk. Our findings suggest that journalists' and stakeholders' attributions of material CSR concerns are negatively associated with stock returns and firms' future earnings surprises.

"In the world of business, bad news often surfaces serially: you see a cockroach in your kitchen; as the days go by, you meet his relatives" - Warren Buffett (2015)

7.1 Introduction

In recent years a number of high-profile companies have been plagued by media allegations regarding poor corporate social responsibility (CSR) practices (Lange and Washburn 2012). Examples range from the misstatement of carbon emissions in Volkswagen, the E.coli scare in Chipotle Inc, defective airbags in Toyota, faulty ignition switches in General Motors Co., to benchmark rigging in investment banks. In each case, media allegations of corporate accidents soon escalated into deeper concerns of negligence. The economic effects of corporate irresponsible behavior have been shown to harm a firm's sales (Mangold and Faulds 2009), disrupt a firm's operations (Hendricks and Singhal 2003), and threaten its license to operate (Suchman 1995), irrespective of whether the media allegations are founded or unfounded (Boydstun et al. 2014; Vanhamme and Grobben 2008; Blackshaw and Nazzaro 2006). Despite the detrimental effects of such media allegations, the extant financial literature has yet to agree on an objective methodology to evaluate the financial materiality of CSR news for at least three reasons. First, the intangible nature of CSR has generated much controversy regarding its measurement as a construct. Prior organizational literature either relies upon measures that lack sufficient depth or contain substantial measurement errors (Waddock and Graves 1997; Daines et al. 2010). Second, investors' abilities to collect CSR information have been hampered by the voluntary nature of sustainability reporting standards. This has resulted in the publication of inconsistent, stale and incomplete information across firms. Third, firms' motivations for publishing such disclosures are often unclear. Recent evidence suggests that firms publish CSR reports merely for symbolic purposes to bolster their social images with consumers (Marquis and Toffel 2012; McDonnell and King 2013) rather than to increase transparency and accountability to investors (Moniz and de Jong 2015). To address the deficiencies associated with the analysis of CSR disclosures, we employ a computational linguistics technique to evaluate the financial materiality of CSR issues from media news.

We retrieve 105,983 news articles from newswires, newspapers, blogs and magazines over the period 1980-2014 and employ a textual analysis to infer journalists' and other stakeholders' perceptions of corporate irresponsible behavior. To infer perceptions we employ a probabilistic topic model adopted from the fields of Natural Language Processing and Information Retrieval. Topic models are statistical models that posit low-dimensional representations of data. The most common topic modeling approach is Latent Dirichlet Allocation (LDA) (Blei et al. 2003). LDA represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words. The goal of topic modeling is to automatically discover the latent topics from a collection of documents. While the documents are observed, the topic structure (the topics, per-document topic distributions, and the per-document per-word topic assignments) are hidden structure (Blei et al. 2003). The output of the model is a probability assignment reflecting the degree to which journalists and other stakeholders express attributions of corporate irresponsible behavior in text. These attributions range from criticisms and concerns over poor CSR practices (without attributions of tangible impacts for a firm) to more material concerns reflecting corporate difficulties and litigation risk. Linguistically, these attributions act as ‘valience shifters’ by serving to accentuate or diminish the underlying sentiment of text (Polanyi and Zaenen 2004; Kennedy and Inkpen 2006). Our findings suggest that journalists’ and stakeholders’ attributions of material CSR concerns are negatively associated with future stock returns and earnings surprises. This finding remains robust after controlling for more traditional measures of media sentiment employed by the extant financial literature (Tetlock 2007, 2008; Loughran and McDonald 2011).

We provide two important contributions to the literature. First, we contribute to the CSR literature by providing an objective methodology to evaluate the financial materiality of CSR news. The analysis of CSR poses a number of challenges largely because of the greater variety and subjectivity of issues compared to accounting information (see Engelberg 2008; Petersen 2004; Demers and Vega 2010). In particular, the inability of researchers to classify intangible information into distinct topics has limited the scope of event studies which assume statistical inference based upon grouping news according to the similarity of the underlying events (see MacKinlay 1997). Consequently, extant CSR studies are typically limited to the analysis of relatively tangible news announcements such as product recalls (Pfarrer et al. 2010), supply chain disruptions (Hendricks and Singhal 2003), environmental pollution fines (Shane and Spicer 1983; Russo and Fouts 1997), and corruption/bribery scandals (Murphy et al. 2009). By contrast, our methodology allows for a finer-grained analysis to classify news along a continuum of intangible and tangible information (see Engelberg 2008; Petersen 2004).

Second, our research contributes to the growing body of textual analysis studies which seek to infer intangible information for the prediction of firms’ earnings. To date, prior studies have retrieved earnings-related information from 10-K reports (Loughran and McDonald 2011; Li 2006; Bao and Datta 2014; Jegadeesh and Wu 2013), conference calls (Price et al. 2012; Mayew and Venkatachalam 2011; Ball et al. 2013; Huang et al. 2014) and press releases (Solomon 2012). These studies suggest that intangible information is costly to process compared to ‘harder’ accounting news. One implication is that intangible information may not be immediately incorporated into stock prices (Engelberg 2008; Demers and Vega 2008). Our approach is intended to encourage further research into the analysis of intangible information for non-financial corpora of texts.

The rest of this study proceeds as follows. Section 7.2 provides a brief review of related research on textual analysis of qualitative information. Section 7.3 outlines the sample data set, variable measurement, and the application of LDA to infer stakeholder attributions. Section 7.4 discusses the main results. Section 7.5 concludes.

7.2 Literature review

7.2.1 Stakeholder theory

Stakeholders are commonly defined as *"any group or individual who can affect or is affected by the achievement of the organization's objectives"*. Primary stakeholders are those groups without whose continuing participation the corporation cannot survive as a going concern and include the firm's customers, suppliers and shareholders (Clarkson 1995; Orlitzky and Benjamin 2001), while secondary stakeholders refer to those groups who influence or affect the firm but who are not engaged in transactions with the firm and are not essential for its survival (such as journalists, regulators and non-governmental organizations). Within the field of organizational studies, stakeholder theory suggests that effective management of stakeholder relationships can mitigate the likelihood of negative regulatory and legislative action (Freeman 1984; Berman et al. 1999), and lead to better firm performance by protecting and enhancing corporate reputation (Fombrun and Shanley 1990; Fombrun 2005; Freeman et al. 2007). Consequently, a company's decision rule to engage in an act of irresponsible behavior depends upon whether its management perceives:

$$\begin{aligned}\text{Private benefit to company} &> \text{E(Reputational cost)} + \text{E(Penalty)} \\ &= \sum p_i \text{ x } RC_i \,|\text{i learns of the allegation} + \pi P\end{aligned} \tag{7.1}$$

where p_i is the probability that stakeholder group i (e.g. consumers, journalists, NGOs, regulators) learns of the misdemeanor, RC_i is the reputational cost associated with the misdemeanor, π is the likelihood of a regulatory penalty and P is the magnitude of such a penalty (see Dyck et al. 2008; Becker 1968).

By publishing CSR news the media can alter p_i, the probability that the firm's behavior is known to a given stakeholder group. The impact of the news media is greater when the news reaches a larger number of stakeholder groups and is published by a salient and credible media source. The second way in which the media may impact a firm's decision rule is via the perceived reputational cost, RC_i, of the news. Empirical studies suggest that the media sets the public agenda by influencing stakeholders' information sets based upon the choice of news topics and their tone influencing how stakeholders' may think about a firm (see McCombs and Shaw 1972; Deephouse 2000). In particular, the media influence stakeholders' impressions of a firm (Pollock and Rindova 2003) by choosing which corporate issues to cover (Kilbanoff et al. 1998; Huberman and Regev 2001; McCombs and Shaw 1972) and how to frame them (Gurun et al. 2012; Gentzkow and Shapiro 2006; Groseclose and Milyo 2005; Dougal et al. 2012). Attribution Theory (McCombs and Shaw 1972), developed in the social psychology literature, addresses how observers form causal inferences and moral judgments to explain irresponsible behavior. Social psychology literature suggests that due to cognitive limitations negative information generates negative human emotional responses and creates cognitive dissonance by altering individuals' views (Burgoon et al. 2002; Festinger 1957). When confronted with negative behavior, individuals spend more time thinking about the issue than positive or neutral behavior, and in turn, search for causal information to explain their behavior (cf. Fiske and Taylor 2008). Finally, the news media may influence the magnitude of a regulatory fine imposed on a firm, either by drawing a regulator's attention to the firm's irresponsible behavior (π) or by similarly influencing the way the regulator thinks about the issue (P) (see also Bednar 2012; Miller 2006).

7.2.2 Textual analysis of intangible information

Prior textual analysis studies conducted in the field of finance typically count the fraction of predefined positive and negative terms in text to infer a composite measure of media sentiment, otherwise known as a 'term counting' approach (seeTetlock 2007, 2008; Loughran and McDonald 2011). The two generally accepted lexicons employed to count terms are the General Inquirer Harvard IV-4 psychosocial dictionary (Stone 1966), referred to hereafter as H4N, and the LM neg word list (Loughran and McDonald 2011). One drawback of term counting is its one dimensional measurement of sentiment based upon detecting 'surface-level' features in text. The approach does not seek to infer contextual information or semantic meaning. To illustrate this, consider the following sets of terms: {'accident', 'unintentional', 'acquitted', 'exonerated'} and {'guilty', 'negligent', 'prosecuted', 'punished'}. These two sets of terms convey the opposite meaning yet each of these terms appear in the H4N and LM word lists with negative polarity. The implication is that each negative term contains equally negative information for a firms' future operating performance. By contrast, our methodology distinguishes between terms and seeks to understand differences in meaning. For example, consider the following news story published in the Dow Jones Newswires corpus (see Tetlock et al. 2008):

"Consumer Groups Say Microsoft Has Overcharged for Software - A study by consumer groups calculates that Microsoft Corp. (MSFT) has overcharged buyers of its software by $10 billion worldwide in the past three years. The alleged pricing abuse will only get worse if Microsoft is not disciplined sternly by the antitrust court".

(Published: January 8, 1999)

Now consider the following news story also published during same month:

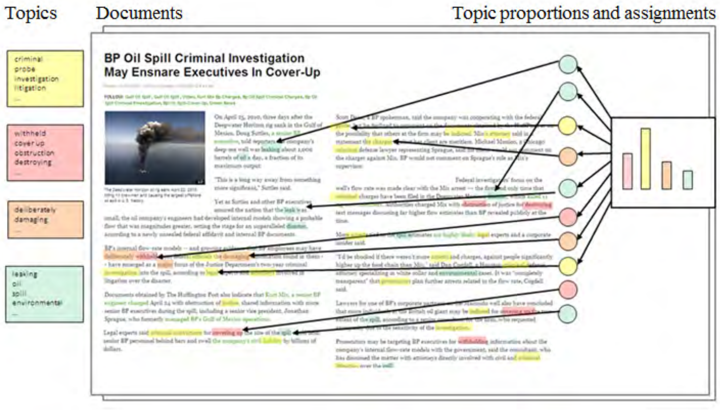
"Environmentalist Grp Claims Bomb Threats Against McDonalds - Accusing McDonalds Corp. of cruelty to animals and other foul practices a militant environmentalist group claimed responsibility for a string of bomb threats".

Both stories exhibit the same degree of media pessimism (measured by the standardized fraction of negative words in the text), yet the cause and nature of the underlying news events differ substantially. The first story suggests that the company knowingly and repeated violated laws while in the second story the company appears to have been victimized.

In contrast to term counting measures of sentiment, LDA employs a finer-grained topic analysis to infer "semantically related probabilistic clusters of words" (see Titov and McDonald 2008). LDA is based on the hypothesis that an author writing news has certain topics in mind. To write about a topic means to pick a word with a certain probability from the pool of words of that topic. A whole document can then be represented as a mixture of different topics. Figure 7.1 provides an extract of a news article to illustrate the methodology. The article discusses both the CSR event, referencing the terms to 'oil', 'spill and 'leak', as well as perceived implications by referring terms including 'criminal' and 'litigation'.

Figure 7.1 Illustrative example of LDA for the classification of CSR news topics

This figure has been adapted from (Blei 2012) and is intended to illustrate the premise of probabilistic topic modelling. LDA assumes that a number of topics which are distributions over words exist for the whole collection (far left). Each document is assumed to be generated as follows: First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored circles) and choose the word from the corresponding topic.



7.3 Sample and variable measurement

7.3.1 Sample construction

We retrieve news articles from the Dow Jones Newswires (DJNS) corpus for the period 1980-2014. DJNS provides the headline, body text, date and timestamp for each news article sourced from on-line newspapers (e.g. The Wall Street Journal), financial blogs and financial magazines. We restrict our search to companies listed in the S&P500 index. We select this group of companies because prior studies suggest that they are more likely to receive media attention and be subjected to more opinionated news (Ahern and Sosyura 2014; Barber and Odean 2008). Following Tetlock et al. 2008, we exclude stories published in the first week after a firm has been added to the S&P index to avoid the documented price momentum effect from influencing the analysis (Shleifer 1986). We also assume that investors do not have time to react to stories that occur after 3:30 pm (30 minutes prior to the market closing), and roll forward the news to the next business day.

We supplement our corpus with four additional data sources. Compustat provides accounting information and earnings announcement days. The Center for Research in Securities Prices (CRSP) reports prices and returns. The Institutional Brokers Estimate System (I/B/E/S) supplies analyst forecast data. Institutional ownership data are obtained from Thomson Financial. Firms included in the sample must have a book value in Compustat, a market value in CRSP at the end of the previous calendar year, and at least one analyst with an earnings estimate no later than 50 days before a company's earnings announcement date. Following prior work, we exclude stocks with prices below $3 to ensure that results are not driven by small, illiquid stocks or bid-ask bounce.

7.3.2 Retrieving company relevant information

The first stage of automated news retrieval requires an evaluation of the relevance of CSR news for a given company. To determine the relevance of news we employ a multinomial Naïve Bayesian text classifier which counts the frequency of company mentions in each article (see Appendix I for details). Antweiler and Frank (2004) employ a similar methodology to detect company mentions in financial media texts. The classifier seeks to distinguish between companies which are the subject of an article versus companies referenced in passing (for example, in the last sentence of the text). The relevance of news is classified based upon detecting the frequency of company mentions in the text. First we remove the suffixes: CL A, CLB, ADR, CO, CORP, HLDG, INC, IND, LTD, and MFG before searching for company names in text (see Engelberg 2008). The algorithm then seeks to detect ticker symbols, official company names, abbreviated names (for example Southwest Airlines is typically referred to as Southwest, while International Business Machines is typically referred to as IBM) and popularized names (e.g. AMR Corp is referred to by its subsidiary American Airlines). Variants of company names are obtained from companies' websites and Wikipedia.

Next we filter news articles to eliminate 'hard' news. Following Antweiler and Frank (2006), we classify articles by detecting keywords in text. Articles with the terms *"earn"*, *"sales", "profit"* or *"eps"* are classified as earnings-related news[1], articles with the terms *"dividend"*, *"split"*, *"offering"*, *"underwriting"* are classified as corporate actions news, and articles with the terms *"acquire"* , *"purchased"*, *"takeover"* as restructuring news. We eliminate these three categories to limit articles to 'softer' corporate news. By construction, the remaining articles in the news corpus consist of legal, environmental, social and corporate governance news. Following Tetlock et al. 2008, we eliminate stories with tables containing quantitative information and impose the requirement that each story must contain at least 50 words in total, with at least 3 words must be listed in either the H4N or LM neg word lists. These requirements are designed to limit the influence of outliers on our sentiment measures.

Finally, we restrict the corpus to 'new' news by filtering out semantically similar articles for the same company. We compute cosine similarity which measures how similar two documents are likely to be in terms of their subject matter. This metric is frequently used in Information Retrieval (IR) research to evaluate the novelty of news (see Manning and Schütze 1999). Documents are characterized by vectors of the frequency terms appear in each document (V_A and V_B respectively):

$$similarity = \cos(\theta) = \frac{V_A . V_B}{\|V_A\| . \|V_B\|} = \frac{\sum_{i=1}^{n} V_{A_i} \ x \ V_{B_i}}{\sqrt{\sum_{i=1}^{n}(V_{A_i})^2} \ x \ \sqrt{\sum_{i=1}^{n}(V_{B_i})^2}} \tag{7.2}$$

where Θ is the angle between the two vectors. The cosine similarity captures the uncentered correlation between two vectors. Its values range from 0 and 1 where greater values indicate greater similarity. Following the IR convention, we retain the first occurrence of a news story in a sequence of related news articles for a given company. We exclude all subsequent stories with a cosine similarity value greater than 0.90 (following Chowdhury et al. 2002). The final corpus consists of 105,983 unique CSR news stories. Table 7.1 reports descriptive statistics for several firm characteristics. We compute Tobin's Q as a measure of firm value, defined as the market value of the firm divided by the replacement value of the firm's assets. The market value of assets is measured as the sum of the book

[1] Searches are conducted on documents in lemmatized form. For example, a search for "earn" captures the terms: "earn" "earns", "earned", "earners", "earning", "earnings" and "earnings-per-share".

value of assets and the market value of common stock outstanding minus the sum of the book value of common stock and balance sheet deferred taxes. Replacement value is represented by the book value of assets (Kaplan and Zingales 1997). Size is the natural log of market capitalization and Book-to-market is book value of equity scaled by market value of equity, both measured as at the previous fiscal year end.

Table 7.1 Descriptive statistics of firm characteristics

This table reports descriptive statistics for several firm characteristics (see text for details). The unit of observation is the firm-year. Sample period: 1980-2014.

	Min.	25th percentile	Median	Mean	75th percentile	Max.
Size	31.30	6662.00	17940.00	55390.00	59760.00	596500.00
Institutional ownership	0.00	0.49	0.62	0.62	0.74	1.00
Analyst coverage	0.00	5.00	13.00	12.98	19.00	44.00
Book-to-market	-5.91	-1.62	-1.05	-1.12	-0.55	2.04
Tobin's Q	0.05	0.78	1.17	2.16	2.30	3.42
Goodwill to assets	0.00	0.01	0.04	0.10	0.14	0.19
Advertising revenue to assets	0.00	0.01	0.02	0.03	0.03	0.05

7.3.3 Inferring the financial materiality of news

We compare three measures to infer journalists' and other stakeholders' attributions of financial materiality. The first measure computes a measure of media pessimism (following Tetlock 2008) and is based upon counting the standardized fraction of negative words in each news article:

$$media\_pessimism = \frac{Neg - \mu_{Neg}}{\sigma_{Neg}} \tag{7.3}$$

where Neg is the number of negative words matched in the H4N dictionary as a fraction of total words in the document, μ_{Neg} is the mean of Neg and σ_{Neg} is the standard deviation of Neg over the prior calendar year. The standardization accounts for potential non-stationarity in the distribution of words as the media changes its coverage or style over time. Terms are stemmed using the Porter algorithm to address the limitation that the H4N word list does not fully account for inflections. For instance, the unigram 'violat' encapsulates the terms 'violate', 'violates', 'violating', 'violated', 'violations'.

Our second measure of media sentiment follows Loughran and McDonald's (2011) methodology. The approach is based on the premise that many terms which appear in the H4N word list are not negative in a financial context (e.g. 'depreciation', 'liability', and 'foreign'). Loughran and McDonald (2011) create a domain specific list of words and employ an alternative weighting scheme to discriminate between words in documents. The term frequency-inverse document frequency (*tf-idf*) weighting scheme, which seeks to scale down frequently occurring terms and scale up rare terms, is commonly used in IR research (see Manning et al. 2008):

$$w_{i,j} = \begin{cases} \dfrac{\left(1+\log\left(tf_{i,j}\right)\right)}{\left(1+\log\left(a_j\right)\right)} log \dfrac{N}{df_i} & if\ tf_{i,j} \geq 1 \\ 0 & otherwise \end{cases} \tag{7.4}$$

where N represents the total number of news stories in the sample, df_i the number of documents containing at least one occurrence of the i^{th} word from the LM neg word list, $tf_{i,j}$ the raw count of the i^{th} word in the j^{th} document, and a_j the average word count in the document.

While tf-idf weighting is effective at discriminating between terms in a document, it also suffers from at least two drawbacks. First, the model makes the simplifying assumption that each document comprises of a single topic. While this assumption may be appropriate for the classification of accounting information in the context of 'earnings' news (see Tetlock 2008; Loughran and McDonald

2011), we start from the premise that the one topic assumption is less likely to be valid for the classification of relatively intangible CSR information due to the greater variety of topics discussed in text (Gurun et al. 2012). Second, tf-idf weighting reveals little about the internal statistical structure of text and makes no use of semantic similarities between words to address the linguistic nuances of synonymy (a keyword may not appear in a document even though the document is closely related to a topic) and polysemy (a keyword may have different meanings in different contexts).

Our third measure employs a probabilistic topic model to discriminate between terms in text. We employ a variant of the standard implementation of LDA to provide a direct comparison with the more traditional measures of media sentiment (Tetlock 2008; Loughran and McDonald 2011). The output of the LDA model are three sets of topic clusters reflecting authors' negative attributions expressed in text, together with probabilistic topic assignments for each document. In the standard implementation of LDA, the topic clusters are unlabelled.We follow a convention from Information Retrieval literature to automatically label the inferred topic clusters (see Lau et al. 2011). Further details are provided in Appendix II. The top terms for each topic cluster and the probabilities of words associated with each topic are displayed in Table 7.2.

Table 7.2 Journalist and stakeholder attributions inferred from the CSR news corpus

This table reports the top terms for each topic cluster and their associated probabilities inferred using the Latent Dirichlet Allocation (LDA) algorithm (Blei et al. 2003). In LDA, a topic is modeled as a probability distribution over a set of words represented by a vocabulary and a document as a probability distribution over a set of topics. Topic labels are generated automatically (see Appendix II for details).

'allegations' topic cluster		'difficulties' topic cluster		'lawsuit' topic cluster	
word	prob.	word	prob.	word	prob.
allege	0.19	difficulties	0.19	illegal	0.17
allegedly	0.17	damaged	0.16	litigations	0.16
complaints	0.13	problematic	0.14	sues	0.13
questionable	0.11	hurting	0.13	failures	0.13
dispute	0.11	negatively	0.12	damaging	0.12
challenged	0.09	critically	0.10	losses	0.11
concerns	0.08	declines	0.07	failing	0.09

The topic clusters appear to reflect differing degrees of negative attributions expressed in text. One topic cluster appears to reflect CSR allegations (complaints, concerns, questionable behavior), a second cluster appears to capture journalists' and stakeholders' attributions of corporate difficulties (with references to damage, problems, declines), while the third topic cluster appears to capture lawsuits (illegal behavior and litigation).

7.4 Results and Discussion

In this section we investigate whether authors' attributions inferred from text are informative for the prediction of firms' earnings. Our premise is that more tangible attributions (e.g. corporate difficulties, litigation) are more likely to be detrimental to a firm versus that CSR allegations that lack attributions of materiality. Following Tetlock (2008), our main test computes each firm's standardized unexpected earnings (SUE). To model firms' earnings, we employ a seasonal random walk model with trend (see Bernard and Thomas 1989):

$$UE_t = E_t - E_{t-4}$$

$$SUE_t = \frac{UE_t - \mu_{UEt}}{\sigma_{UEt}} \qquad \textbf{(7.5)}$$

where E_t is the firm's earnings in quarter t, and the trend and volatility of unexpected earnings (UE) are equal to the mean (μ) and standard deviation (σ) of the firm's previous 20 quarters of unexpected earnings data, respectively.

We require that each firm has non-missing earnings data for the most recent 10 quarters and assume a zero trend for all firms with fewer than 4 years of earnings data. We use the median analyst forecast from the most recent period in the I/B/E/S summary file prior to the earnings announcement. We winsorize SUE and all analyst forecast variables at the 1% level to reduce the impact of estimation error and extreme outliers, respectively. To align different frequencies of data, we create a composite document for each firm by aggregating news articles between the consecutive earnings announcement dates of firms. We include control variables based on a firm's lagged earnings, size, book-to-market ratio, analysts' earnings forecast revisions, and analysts' forecast dispersion. We measure firms' lagged earnings using the last quarter's SUE. We compute analysts' forecast dispersion (Forecast Dispersion) as the standard deviation of analysts' earnings forecasts in the most recent time period prior to the announcement scaled by earnings volatility (σ). Finally, we control for a firm's cumulative abnormal return, Pre_FFAlpha, estimated from the intercept of the event study regression. A Fama-French three-factor model is estimated using daily returns between days -252 to -31 prior to the news release. The regression results are reported in Table 7.3.

Table 7.3 Regressions of journalist and stakeholder attributions and earnings surprises
The dependent variable in each regression is the standardized unexpected earnings (SUE). Media pessimism (Tetlock) refers to the term counting measure employed by Tetlock (2008) and uses the H4N word list. Media pessimism (LM) measure refers to the tf-idf weighted values employed by Loughran and McDonald (2011) and uses the LM neg-list. Allegations captures stakeholder perceptions of CSR allegations (e.g. complaints, concerns, questionable behavior) as inferred by the LDA model. Difficulties captures stakeholder perceptions of corporate difficulties (e.g. references to damage, problems, declines) as inferred by the LDA model. Lawsuits captures stakeholder perceptions of CSR litigation risk as inferred by the LDA model. Pre_FFAlpha is a firm's cumulative abnormal return estimated from the intercept of the event study regression over the [–252,–31] time window. Fama-French (1997) industry dummies (based on 48 industries) and a constant are also included in each regression. Standard errors are adjusted for clustering effects following Petersen (2009).

	(1)	(2)
Media pessimism (Tetlock)	-0.0539	
	(-1.38)	
Media pessimism (LM)		-0.0546
		(-1.43)
Allegations	0.1498	0.2091
	(1.38)	(1.53)
Difficulties	-0.1536	-0.1996
	(-1.98)	(-2.08)
Lawsuit	-0.2600	-0.3607
	(2.08)	(-2.26)
Forecast dispersion	0.0014	0.0017
	(1.48)	(1.50)
Forecast Revisions	8.1404	8.1404
	(10.53)	(10.52)
Log(Market Equity)	-0.0224	-0.0255
	(-4.91)	(-5.27)
Log(Book/Market)	-0.1503	-0.1592
	(-5.07)	(-5.12)
Pre_FFAlpha	-0.1100	-0.1140
	(-0.51)	(-0.50)
Adjusted R^2	0.11	0.12

Column (1) presents the results controlling for the media pessimism measure employed by Tetlock et al. (2008) based upon the H4N word list, while column (2) controls for tf-idf weighted measure employed by Loughran and McDonald (2011) based upon the LM neg-list. In both regressions, we find that the traditional 'term counting' measures of sentiment are statistically insignificant. By contrast, we find a statistically significant relation between journalists' and stakeholders' attributions of tangible negative news (either corporate difficulties or litigation risk) and subsequent earnings surprises. Linguistically, this finding is consistent with the view that authors' attributions act as ‘valience shifters’ which serve to accentuate or diminish the underlying sentiment of a document (see also Marcus and Goodman 1991). Taken together, our findings suggest that finer-grained linguistic measures can more accurately infer intangible information for the prediction of firms’ earnings.

Next we estimate an event-study regression and test the hypothesis that journalists’ and stakeholders’ attributions are informative for stock returns. We estimate abnormal returns over a CRSP value-weighted benchmark using the Fama-French (1993) three-factor model with an estimation window of [–252,–31] trading days prior to the news announcement (see Tetlock 2008). The regressions include control variables capturing firm size and the book-to-market ratio. We measure firm size (Log(Market Equity)) and book-to-market (Log(Book/Market)) at the end of the preceding calendar year, following Fama and French (1992). We estimate a pooled ordinary least squares regression with robust standard errors. The regression results are reported in Table 7.4.

Table 7.4 Regressions of journalist and stakeholder attributions and stock returns
The dependent variable in each regression is the event period excess return (defined as the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return over the three-day event window, expressed as a percent). The Media pessimism (Tetlock) measure refers to the term counting measure employed by Tetlock (2008) and uses the H4N word list. The Media pessimism (LM) measure refers to the tf-idf weighted values employed by Loughran and McDonald (2011) and uses the LM neg-list. Allegations captures stakeholder perceptions of CSR allegations (e.g. complaints, concerns, questionable behavior) as inferred by the LDA model. Difficulties captures stakeholder perceptions of corporate difficulties (e.g. references to damage, problems, declines) as inferred by the LDA model. Lawsuits captures stakeholder perceptions of CSR litigation risk as inferred by the LDA model. Pre_FFAlpha is a firm's cumulative abnormal return estimated from the intercept of the event study regression over the [–252,–31] time window. Fama-French (1997) industry dummies (based on 48 industries) and a constant are also included in each regression. Standard errors are adjusted for clustering effects following Petersen (2009).

	(1)	(2)
Media pessimism (Tetlock)	-0.0046	
	(1.058)	
Media pessimism (LM)		-0.0015
		(-1.82)
Allegations	0.0064	0.0020
	(1.38)	(1.53)
Difficulties	-0.0315	-0.0270
	(-2.11)	(-2.26)
Lawsuit	-0.0739	-0.0821
	(-2.42)	(-2.91)
Log(Market Equity)	-0.0032	-0.0029
	(-4.36)	(-4.04)
Log(Book/Market)	-0.0015	-0.0048
	(-3.86)	(-3.56)
Pre_FFAlpha	-0.0013	-0.0013
	(-1.27)	(-1.29)
Adjusted R^2	0.02	0.02

Column (1) presents the results controlling for the media pessimism measure employed by Tetlock et al. (2008) based upon the H4N word list, while column (2) controls for tf-idf weighted measure employed by Loughran and McDonald (2011) based upon the LM neg-list. In both cases, subsequent abnormal stock returns appear to be negatively correlated with the release of materially negative news (attributions of corporate difficulties and litigation risk).

7.5 Conclusion

In this study, we employ a probabilistic topic model to infer journalists' and stakeholders attributions of corporate irresponsible behavior. The model identifies three distinct clusters of attributions which appear to vary in their degree of financial materiality. Linguistically, the inferred topic clusters appear to act as 'valience shifters' and either serve to accentuate or diminish the underlying sentiment of a document. Our findings suggest that differences in attributions are associated with differences in return predictability and future earnings surprises.

Our findings have important practical implications both for investors and corporate managers. To date, three organizations – the International Integrated Reporting Council (IIRC), the Sustainability Accounting Standards Board (SASB) and the Global Reporting Initiative (GRI) diverge in their approaches to define materiality. This has hindered the ability of corporations and investors to integrate sustainability considerations into their decision making processes. Our approach provides an objective framework to evaluate the financial materiality of CSR news and seeks to address the call from financial literature to classify information along a continuum of intangible and tangible news (Engelberg 2008; Petersen 2004).

7.6 Appendix I

In this section we describe the methodology to detect journalists' and stakeholders' attributions in text. We employ a variant of LDA, an unsupervised, generative model which proposes a stochastic procedure by which words in documents are generated (Blei 2003).

The distinguishing feature of LDA is its ability to model multiple topics in contrast to the Naïve Bayesian classifier which assumes that a document is generated by first choosing only one topic z, and then generating *N* words independently from the conditional multinomial distribution *p(w\z)*:

$$p(w) = \sum\nolimits_{z=1}^{k} \left(\prod\nolimits_{n=1}^{N} p(w_n \setminus z)\right) p(z) \tag{7.6}$$

LDA relaxes the assumption of one topic and represents each document *d* with a *K*-dimensional mixed membership vector Θ^d which sums to one. Given a corpus of unlabeled text documents, the model discovers hidden topics as distributions over the words in the vocabulary. Words are modeled as observed random variables, while topics are observed as latent random variables. Once the generative procedure is established, we may define its joint distribution and then use statistical inference to compute the probability distribution over the latent variables, conditioned on the observed variables. The LDA model assumes that there are k underlying latent topics according to which documents are generated, and that each topic is represented as a multinomial distribution over the |V| words in the vocabulary.

More formally:

A document of N words w = $\{w_1,\ldots,w_N\}$ is generated by the following process. First Θ is sampled from a Dirichlet $(\alpha_1,\ldots, \alpha_k)$ distribution such that $\Theta_i \geq 0$, $\Sigma_i\ \Theta_i=1$. For each of the N words, a topic z_n $\{1,\ldots k\}$ is sampled from a Mult(Θ) distribution $p(z_n = i\setminus\Theta) = \Theta_i$. The multinomial distribution is chosen because the model relies on the computation of discrete counts of co-occurrences of words in documents, and that the Dirichlet distribution is chosen because it is conjugate to the multinomial distribution, which makes the computations for inference and parameter estimation easier.

Each word w_n is sampled, conditioned on the z_n^{th} topic, from the multinomial distribution $p(w\setminus z_n)$. The probability of a document is the following mixture:

$$p(w) = \int\limits_{\theta} \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n\, | z_n;\ \beta) p(z_n|\theta)\right) p(\theta; \alpha)\, d\theta \tag{7.7}$$

Where $p(\Theta;\alpha)$ is Dirichlet, $p(z_n \backslash \Theta)$ is a multinomial parameterized by Θ, and $p(w_n \backslash z_n;\ \beta)$ is a multinomial over the words. The model is parameterized by the k-dimensional Dirichlet parameters $\alpha = (\alpha_1, \ldots, \alpha_k)$ and a k x |V| matrix β, which are parameters controlling the k multinomial distributions over words. The computational problem is the calculation of conditional distribution of the topic structure given the observed documents. The posterior is:

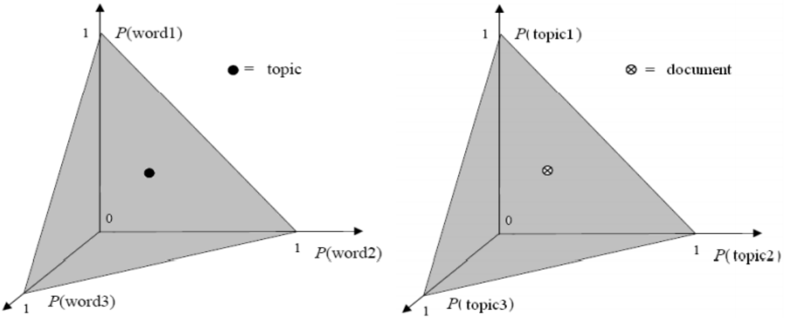
$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \backslash\, w_{1:D}\,) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{7.8}$$

The numerator is the joint distribution of all the random variables. The denominator is the marginal probability of the observations, which is the probability of seeing the observed corpus under any topic model and is estimated by Gibbs sampling. The goal of topic modeling is then to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure (the topics, per-document topic distributions, and the per-document per-word topic assignments) are hidden. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure (Blei et al. 2003).

The probabilistic topic model has an elegant geometric interpretation of the relations between document-topic and topic-word as shown in Figure 7.2. With a vocabulary containing N distinct word types, a N dimensional space can be constructed where each axis represents the probability of observing a particular word type. The N-1 dimensional simplex represents all probability distributions over words. Consequently, topics and documents (convex combination of the K topics) are represented as points on the simplex.

Figure 7.2: Geometric interpretation of LDA

The illustration below is sourced from Qin et al. (2009). A topic is a distribution over observable words and a document is a distribution of latent topics. A document can be represented as a point on the surface of a simplex of topics. A topic can be regarded as a point of on the simplex of words. In this simple case, there are 3 words and 3 topics.



The Dirichlet prior on the topic-word distributions can be interpreted as forces on the topic locations. Higher values of β move the topic locations away from the corners of the simplex. Thus when the number of topics is smaller than the number of words, the projection of each document onto the low-dimensional subsimplex can be thought of as dimension reduction (Steyvers and Griffiths 2006).

7.7 Appendix II

We employ a variant of the standard implementation of LDA to guide the topic model by providing a set of seed words of interest. Prior NLP studies show that seed words can help guide the topic discovery process (see Jagarlamudi et al. 2012). The approach is designed to improve both topic word distributions (by biasing topics to produce appropriate seed words) and to improve document-topic distributions (by biasing documents to select topics). Specifically, we provide the negative terms listed in the H4N and LM neg word lists as seed words in the LDA model. We implement standard settings for LDA hyperparameters with $\alpha = 50/K$ and β=.01 following (Griffiths and Steyvers 2004). The number of topics, K, is inferred by maximizing the likelihood of fitting the LDA model over the corpus of documents.

The output of the LDA model is a set of topic clusters reflecting negative attributions expressed in text. To label the LDA topic clusters, we employ an automated approach similar to Lau et al. (2011). For each inferred topic cluster (z), the top five terms (w) ranked by their marginal probabilities $p(w|z)$ are entered as keywords into an online search query. The approach is intended to retrieve a wider selection of online documents to reflect the topic cluster.

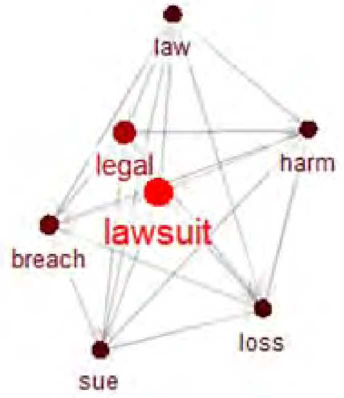
Next we retrieve the titles and summary texts from the search engine query results and employ a link analysis to detect the most frequent terms mentioned in the text. Graph-based algorithms have received much attention (Mihalcea and Tarau 2004) as an approach to keyphrase extraction and are considered to be state-of-the-art unsupervised methods (Liu et al. 2009). In a graph representation of a document, nodes are words or phrases, and edges represent co-occurrence or semantic relations. The underlying assumption is that all words in the text have some relationship to all other words in the text. Such an approach is statistical, because it links all co-occurring terms without considering their meaning or function in text. Centrality is often used to estimate the importance of a word in a document (Opsahl et al. 2010), and is a way of deciding on the importance of a vertex within a graph that takes into account global information recursively computed from the entire graph, rather than relying only on local vertex-specific information (Boudin 2013). Specifically, we employ TextRank (Mihalcea and Tarau 2004), a ranking algorithm based on the concept of eigenvector centrality, computes the importance of the nodes in the graph. Each vertex corresponds to a word. A weight, w_{ij}, is assigned to the edge connecting the two vertices, v_i and v_j. The goal is to compute the score of each vertex, which reflects its importance, and use the word types that correspond to the highest scored vertices to form keywords for the text (Boudin 2013). The score for v_i, $S(v_i)$, is initialized with a default value and is computed in an iterative manner until convergence using recursive formula shown in Equation (7.9).

$$S(v_i) = (1 - d) + d\ x \sum_{v_j \in Adj(v_i)} \frac{w_{ji}}{\sum_{v_k \in Adj(v_j)} w_{jk}} S(v_j) \qquad \textbf{(7.9)}$$

where Adj(v_i) denotes v_i's neighbors and d is the damping factor set to 0.85 (Mihalcea and Tarau 2004). Figure 7.3 displays the resulting clustering of terms. The size of each node is directly proportional to the TextRank score.

Figure 7.3: Link analysis approach used to generate seed words for the LDA model

This figure displays the results of a link analysis based on detecting the most frequently mentioned terms generated from an online search query for terms in one particular topic cluster. TextRank (Mihalcea and Tarau 2004) a ranking algorithm based on the concept of eigenvector centrality, is employed to compute the importance of the nodes in the graph. The size of each node is directly proportional to the TextRank score. The algorithm determines 'lawsuit' as the most frequently occurring term. This term is used to label the LDA topic cluster. Different nodes colors reflect different communities identified using the Clauset-Newman-Moore algorithm and are for illustrative purposes only.



Chapter 8

"We are living in the midst of a social, economic, and technological revolution. How we communicate, socialize, spend leisure time, and conduct business has moved onto the Internet...The resulting explosion of data and discovery is changing our world." (Podesta et al. 2014)

8.1 Conclusion

In this thesis we have presented a series of tools and case studies to infer intangible information using a framework of probabilistic topic models. This need is particularly important because of the conservative nature of US and international accounting standards which limits firms' abilities to recognize most types of intangible assets on their balance sheets.

Our first research question asks:

- **How can we use online texts to infer intangible information for firms?**

The traditional approach to inferring the measurement of intangible information relies upon survey-based measures. In this thesis, we demonstrate the merits of an automated approach to infer stakeholders' and journalists opinions of a firm using unstructured data. To determine whose opinions are relevant, we employ a stakeholder theory perspective and consider a firm's customers, local communities, employees, investors, journalists, and regulators. In particular, stakeholder theory suggests that the effective management of stakeholder relationships can mitigate the likelihood of negative regulatory and legislative action (Freeman 1984) and increase firm performance by protecting and enhancing corporate reputation (Fombrun and Shanley 1990; Fombrun 2005; Freeman et al. 2007). Chapter 2 investigates a shareholder perspective to infer a measure of reputation. Chapter 3 considers society's perspective to infer a measure of firms' environmental sustainability, while Chapter 4 evaluates reputation and credibility from a regulatory perspective. The goal of these chapters is to demonstrate the merits of LDA to compute finer-grained measures of sentiment compared to the traditional 'term counting' measures employed by the extant financial literature.

Our second research question asks:

- **How can we integrate intangible information for investment analysis?**

Traditionally, investors have relied upon financial statement analysis to make informed investment decisions. In this thesis, we describe a framework to automate the measurement of intangible information for a large number of companies. In particular, we highlight the merits of LDA as a dimension reduction technique to detect specific aspects of a firm's intangible assets (e.g. corporate culture, environmental sustainability) and intangible liabilities (reputational damage). Chapters 5-7 draw upon techniques from financial asset pricing literature. We demonstrate how measures of intangible information can be combined with 'hard' accounting data reported in balance sheets, income and cash-flow statements in a regression framework to predict the future earnings and returns of a company.

Finally, our third research question asks:

- **Is intangible information incremental to the prediction of firms' earnings?**

To evaluate the benefits of our measures we estimate regressions to predict firm's earnings. We start from the premise that if intangible assets (liabilities) cause positive (negative) stock returns because of financial analysts' "errors-in-expectations", then financial analysts' forecasts of future earnings should be systematically too low (high) relative to actual earnings (Edmans 2011).

8.2 Contributions

The primary contribution of this thesis is to highlight the merits of textual analysis to infer the intangible information for a firm and to integrate measures of intangible information into investment decision processes. While the extant financial literature has sought to retrieve intangible information from financial corpora such as the Wall Street Journal (Tetlock 2007, 2008), 10-K regulatory filings (Loughran and MacDonald 2010; Li 2006; Jegadeesh and Wu 2013), and companies' quarterly earnings conference calls (Price et al. 2012; Mayew and Venkatachalam 2011), these studies provide a one-dimensional perspective of text based upon accounting-related topics. Tetlock et al. (2008) models the 'earnings' of a company while Li (2006) evaluates the 'risks'. By contrast, we employ a technique to infer a broad range of intangible information from a variety of non-accounting sources (CSR disclosures, employee blogs and corporate governance news). This is achieved using an automated and objective approach to classify information while taking into account replicability. In particular, LDA is robust to an author's particular choice of words, thereby mitigating the criticism levied by Loughran and McDonald (2011) that if corporate managers know there is a list of words that have a significant negative impact on returns then they will systematically avoid these words going forward.

Second, our research contributes to NLP literature by demonstrating the applications of aspect-level sentiment measures to the financial domain. In recent years, sentiment analysis has become a popular research area in computational linguistics partly because of the explosion of unstructured information freely available on the Web. To the best of our knowledge, prior NLP studies have mostly sought to infer sentiment from the perspective of consumers e.g. from online customer reviews (see Hu and Liu

2004). Given a collection of review texts, the goal is to infer product aspects using star ratings as a 'gold standard' for evaluation. The task of consumer review sentiment analysis usually involves techniques to process texts which may be limited in length, potentially with misspellings, colloquialisms, shortened forms of words and/or emoticons. In this thesis, we argue that sentiment analysis within the financial domain imposes a different set of challenges. First we highlight the importance of inferring opinions across multiple stakeholder groups (including consumers, employees, financial media journalists, and regulators). Second, our approach to aspect detection draws upon financial domain knowledge to steer the LDA model towards topics of interest. We achieve this by providing sets of seed words from financial lexicons (e.g. the Harvard IV-4 and Loughran and McDonald neg word lists) and online databases (e.g. DBPedia and SPARQL endpoints). The seed words improve both topic word distributions (by biasing topics to produce appropriate seed words) and to improve document-topic distributions (by biasing documents to select topics). Third, we demonstrate how to integrate unstructured text into traditional financial databases. In particular, we design algorithms to string match company names in text to official company identifiers contained in the CRSP database. Fourth, in the absence of a 'gold standard' benchmark to evaluate the measurement of intangible information, we assess their relevance for the prediction of firms' earnings and stock returns.

The third contribution of our research is to organizational literature. Fombrun and van Riel (1997) suggest that the lack of attention to the measurement of intangible assets can be traced to the diversity of literatures seeking to measure such constructs, while Mahon (2002) suggests that different disciplines make 'little or no reference to the parallel research being conducted elsewhere'. This thesis seeks to bridge the gap between the literature. Topic modeling provides an objective means to infer stakeholders' perceptions of a company at a higher frequency than the survey-based measures currently employed in the extant literature (see Edmans 2011). In this thesis we argue that surveys are manual and time-consuming to produce, and are thus limited in scope with regards to the number of questions they can ask, the number of companies they can cover and their timeliness to collect and process responses. By contrast, our measures are based on publicly available datasets for a large cross-section of companies, as opposed to more bespoke measures typically used in current research and practice. The results of our analysis have important practical implications for investors. The high costs associated with gathering and processing unstructured data suggests that investors may overlook intangible information compared to more structured financial data (Da et al. 2011).

8.3 Limitations and final thoughts

The limitations of our analysis can be broadly classified into choices regarding datasets, design methodologies, evaluation methods and implementation.

The greatest limitation of unstructured datasets is the potential for sample selection bias. There are numerous reasons why individuals may choose to share their opinions on the Web, each of which can bias the statistical inference of topic and sentiment analysis. In Chapter 3, we draw upon organizational literature which suggests that companies may selectively publish sustainability topics depending upon NGOs' demands. Companies are incentivized to spin information so that they are seen by their stakeholders in a positive light. In Chapter 5 we provide evidence of a selection bias in the media coverage of Chinese corporate governance news. Our findings are consistent with the view that media coverage is a function of profit maximizing incentives rather than an exogenous decision (Bushee et al. 2010; Gentzkow and Shapiro 2006; Mullainathan and Shleifer 2005). One way to limit the impact of these reporting biases is to retrieve a broader set of documents beyond those reported by a company and the news media. This may afford a more holistic view of a company based upon the consideration of multiple stakeholders. Nonetheless, there is still the potential for selection bias due to the difficulties associated with evaluating the credibility of information. In Chapter 6, we argue that employees may be biased by their prior experiences or by cultural differences which may influence their choice of words expressed in text. To mitigate these selection biases we inferred sentiment from specific groups of employees and aggregated information across reviews to limit the impact of any one reviewer influencing the overall perception of a company. Despite these adjustments, we still believe that reviewers' motives for openly sharing such information may bias the aggregated opinions derived from the reviews.

The methodological limitations of our research largely concern the assumptions underpinning the LDA model's generative process. The model's assumptions can impact topic classifications and their semantic coherence (human interpretability). One limitation of LDA is its inability to model correlation between topics. By contrast, a Correlated Topic Model (CTM) builds upon LDA and employs a more flexible distribution for topic proportions by accounting for the covariance structures between topics (Blei and Lafferty 2006a, 2007). This results in a more realistic model when the presence of one latent topic may be correlated with the presence of another. A second limitation of LDA is its failure to capture how topics evolve over time. In a Dynamic Topic Model (DTM), a document is divided into sequential segments (e.g. by year). The DTM model is then applied to the segmented corpus allowing topic distributions to evolve from segment to segment resulting in a hierarchical model of sequential document collections (Blei and Lafferty 2006b, 2009). These extensions to LDA provide promising avenues for research. We stress that the goal of this thesis was to highlight the merits of LDA to infer measures of intangible information rather than to compare topic modeling approaches.

Evaluation concerns largely arise because we do not use direct measures to value the intangible assets and liabilities of a firm. While these limitations suggest that the conclusions from our analysis should be interpreted with some caution, we stress that we are constrained by the availability of such information due to the conservative nature of accounting standards which prevents firms from recording most types of intangible assets in their financial statements (see IAS 38[1]). To overcome this deficiency, we rely upon a statistical approach to evaluate the relation between indirect measures and firms' characteristics (earnings surprises and stock returns). We are cautious to draw conclusions regarding the nature of a causal relationship for the prediction of stock returns due to the potential for model misspecification. In particular, the absence of regression control variables can hinder statistical inference (see Fama and French 1992).

The implementation limitations of our research refer to the feasibility of employing LDA to large-scale document analysis. Fitting a LDA topic model requires approximate inference techniques that are computationally expensive. Large-scale datasets bring significant challenges for machine learning, particularly in terms of computation time and memory requirements. A natural extension of LDA to address these time and memory issues is take advantage of multiprocessor/multicore technology (high performance computing).

8.4 Final thoughts

The Internet has accumulated a huge and growing amount of digital information including news, blogs, web pages, images, audio, video and social networking data. The vast amount of information, commonly referred to as 'big data', is often characterised by the three 'Vs' – 'Volume', 'Veracity' and 'Variety'. The 'volume' of unstructured data has changed the role of information in financial markets. Traditional investment processes cannot possibly perform information retrieval tasks on the ever increasing volume of unstructured text that is being generated. One consequence of this is that intangible information may be partially overlooked by financial analysts because of the high costs associated with searching, retrieving and processing unstructured data. In our view, the exploitation of unstructured data holds promise for investors seeking to integrate a wider variety of information into their investment models beyond traditional financial statement analysis. The 'variety' of data refers to the range of stakeholder perspectives available online. Automated information retrieval approaches can provide investors with a more holistic view to 'look inside' a company compared to more

[1] http://www.ifrs.org/IFRSs/Documents/Technical-summaries-2014/IAS%2038.pdf

traditional financial accounting analysis. Finally, the ‘velocity’ of data makes information usable at much higher frequency, enabling investors to make timely and more informed decisions.

References

Adams, J. S., (1963), Toward an understanding of inequity, *Journal of Abnormal and Social Psychology*, Vol. 67, pp. 422-436.

Ahern, K. R., Sosyura, D., (2014), Who Writes the News? Corporate Press Releases during Merger Negotiations, *Journal of Finance*, Vol. 69, pp. 241-291.

Alm, C. O., (2009), Affect in Text and Speech. VDM Verlag: Saarbrücken.

Ambrose, M. L., Kulik, C. T., (1999), Old friends, new faces: Motivation in the 1990s. *Journal of Management*, Vol. 25, pp. 231–292.

Ang, S. J., Jiang, Z.Q., Wu, C.P., (2012), Good apples, Bad apples: Sorting among Chinese companies traded in the US. Working paper, Florida State University and Xiamen University.

Angelopoulos, G., Giamouridis, D., Vlismas, O., (2012), Inferring the Value of Intangible Assets, Working paper, Edhec Business School.

Antweiler, W., Frank, M. Z., (2004), Is All that Talk just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance*, Vol. 59, pp. 1259–1294.

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

Ashforth, B. E, Gibbs, B.W., (1990), The double-edge of organizational legitimation, *Organization Science*, Vol. 1, No. 2, pp. 177-94.

Bagheri, A., Saraee, M., de Jong, F., (2013a), Care more about customers: unsupervised domain independent aspect detection for sentiment analysis of customer reviews, Knowledge-Based Systems, Vol. 52, pp. 201-213.

Bagheri, A., Saraee, M., de Jong, F., (2013b), An unsupervised aspect detection model for sentiment analysis of reviews, Proceedings of Natural Language Processing and Information Systems, Springer Berlin Heidelberg, pp. 140-151.

Baker, R. R., Biddle, G. C., O'Connor, N., (2012), SOX internal control deficiencies and auditors of U.S. - listed Chinese versus U.S. firms, Working paper, The University of Hong Kong.

Ball, C., Hoberg, G., Maksimovic, V., (2013), Disclosure Informativeness and the Tradeoff Hypothesis: A Text-Based Analysis, University of Maryland Working Paper.

Bank of England, (2013), Monetary policy trade-offs and forward guidance. http://www.bankofengland.co.uk/publications/Documents/inflationreport/2013/ir13augforwardguidance.pdf

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., Yu, Y., (2009), Joint emotion-topic modeling for social affective text mining, ICDM.

Bao, Y., Datta, A., (2014), Simultaneously discovering and quantifying risk types from textual risk disclosures, *Management Science,* Vol 60, No. 6, pp.1371-1391.

Barber, B., Odean, T., (2008), All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies*, Vol. 21, No. 2, pp. 785–818.

Barney, J., (1991), Firm resources and sustained competitive advantage, *Journal of Management*, Vol 17, pp. 99–120.

Barney, J.B., Wright, P.M., (1998), On becoming a strategic partner: The role of human resources in gaining competitive advantage. *Human Resource Management,* Vol. 37, pp. 31- 46.

Baumeister, R., F., Bratslavsky, E., Finkenauer, C., and Vohs, K., D., (2001), Bad is stronger than good, *Review of General Psychology,* Vol. 5, pp. 323–370.

Becker, G., S., (1968), Crime and punishment: An economic approach, *Journal of Political Economy,* Vol. 76, pp. 169-217.

Bednar, M. K., (2012), Watchdog or lapdog: A behavioral view of the media as a governance mechanism, *Academy of Management Journal*, Vol. 55, No. 1.

Berman, S. L., Wicks, A. C., Kotha, S., Jones, T. M., (1999), Does Stakeholder Orientation Matter? The Relationship between Stakeholder Management Models and Firm Financial Performance, *Academy of Management Journal*, Vol 42, No. 5, pp. 488–506.

Bernard, V., Thomas, J., (1989), Post-earnings-announcement drift: Delayed price response or risk premium?, *Journal of Accounting Research*, Suppl. 27, pp. 1-36.

Besanko, D, Dranove, D., Shanley, M., (2000), Economics of Strategy, John Wiley & Sons.

Bholat, D., Hansen, S., Santos, P., Schonhardt-Bailey, C. (2015), Handbook - No. 33 Text mining for central banks, Centre for Central Banking Studies, Bank of England.

Blackshaw, P., Nazzaro, M., (2006), Consumer-generated media (CGM) 101: Word-of-mouth in the age of the web-fortified consumer, New York: Nielsen BuzzMetrics.

Blaug, R., Lekhi, R., (2009), Accounting for intangibles: Financial reporting and value creation in the knowledge economy, A Research Report for The Work Foundation's Knowledge Economy Programme.

Blei, D., (2012), Probabilistic Topic Models. Communications of the ACM, Vol. 55, No. 4.

Blei, D. M., Lafferty, J. D., (2006a), Dynamic topic models, Proceedings of the 23rd international conference on Machine learning, ACM, pp. 113–120.

Blei, D. M., Lafferty, J. D., (2006b), Correlated topic models, Weiss, Y., Schölkopf, B., and Platt, J., editors, Advances in Neural Information Processing Systems 18. MIT Press, Cambridge, MA.

Blei, D.M, Lafferty, J.D., (2007), A correlated topic model of science, The Annals of Applied Statistics, pp. 17–35.

Blei, D., and Lafferty, J., (2009), Topic models, Srivastava, A., and Sahami, M., eds., Text Mining: Theory and Applications. Taylor and Francis

Blei, D., M., Ng, A., Jordan, M., I., (2003), Latent Dirichlet Allocation*, Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Bollen, J., Mao, H. and Zeng, X.J., (2010), Twitter mood predicts the stock market, *Journal of Computational Science*, Vol 2, No. 1, pp. 1–8.

Borgers, A., Derwall, J., Koedijk, K., ter Horst, J., (2013), Stakeholder Relations and Stock Returns: On Errors in Expectations and Learning, *Journal of Empirical Finance*, Vol. 22, No. 1, pp. 159-175.

Boswell, W. R., (2006), Aligning employees with the organization's strategic objectives: out of 'line of sight', out of mind, *International Journal of Human Resource Management*, Vol. 17, pp. 1489-511.

Boswell, W. R., Boudreau, J. W., (2001), How leading companies create, measure, and achieve strategic results through "line of sight", *Management Decision*, Vol. 39, pp. 851-859.

Boudin, F. A, (2013), Comparison of Centrality Measures for Graph-Based Keyphrase Extraction, International Joint Conference on Natural Language Processing (IJCNLP), Oct 2013, pp. 834-838.

Boukus, E., Rosenberg, J.V., (2006), The information content of FOMC minutes. Federal Reserve Bank of New York Working Paper.

Boydstun, A. E., Hardy, A., Walgrave, S., (2014), Two Faces of Media Attention: Media Storm Versus Non-Storm Coverage, *Political Communication*, Vol. 31, No. 4, pp. 509-531.

Bozic, B., Peters-Anders J., Schimak, G., (2014), Ontology Mapping in Semantic Time Series Processing and Climate Change Prediction, 7th Intl. Congress on Env. Modelling.

Breiman, L., (2001), Random Forests. Machine Learning, Vol. 45, No. 1, pp. 5–32.

Breslau, L., Cao, P., Fan, L., Phillips, G., Shenker, S., (1999), Web caching and Zipf-like distributions: Evidence and implications Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies.

Brin, S., Page, L. (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998).

Brockman, P., Salas, J. M., Zagorchev, A., (2015), The Impact of Cross-Listing on Corporate Governance: A Test of the Governance Bonding Hypothesis, Working paper.

Brody, S., Elhadad, N., (2010), An unsupervised aspect-sentiment model for online reviews, Proceeding HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 804-812

Brown, B., Perry, S., (1994), Removing the financial performance halo from Fortune's Most Admired companies, *Academy of Management Journal*, Vol. 37, pp. 1347-1359.

Burgoon, M., Denning, V., P., Roberts, L., (2002), Language expectancy theory, In The persuasion book: Developments in theory and practice. J.P. Dillard & M. Pfau (Eds.), Chapter 7, pp. 117-136.

Bushee, B., Core, J., Guay, W., Hamm, S. J.W., (2010), The Role of the Business Press as an Information Intermediary, *Journal of Accounting Research*, Vol. 48, pp. 1–19.

Carney, M., (2013), Monetary Policy After the Fall, Eric J. Hanson Memorial Lecture, University of Alberta, Edmonton, Alberta.

Carhart, M. M., (1997), On the persistence of mutual fund performance, *Journal of Finance*, Vol. 52, pp. 57-82.

Chan, L., Jegadeesh, N., Lakonishok, J., (1996), Momentum strategies, *Journal of Finance,* Vol. 51, pp. 1681-1713.

Chan, L., Lakonishok, J., Sougiannis, T., (2001), The stock market valuation of research and development expenditures, *Journal of Finance*, Vol. 56, pp. 2431-2456.

Chan, W. S., (2003), Stock price reaction to news and no-news: drift and reversal after headlines, *Journal of Financial Economics*, Vol. 70, pp. 223–260.

Charniak, E., (1997), Statistical techniques for natural language parsing, AI Magazine, 18, 33-43.

Chen, C. C., Meindl, J. R., (1991), The construction of leadership images in the popular press: The case of Donald Burr and People Express, *Administrative Science Quarterly*, Vol. 36, pp. 521−551.

Chen, Y., Lee, S.Y. M., Li, S., and Huang, C.R., (2010), Emotion Cause Detection with Linguistic Constructions. Proceedings of the 23rd International Conference on Computational Linguistics, pp. 179-187.

Chen, K.C., Lin, Y.C., Lin, Y.C., (2012), Does foreign company's shortcut to Wall Street cut short their financial reporting quality? Evidence from Chinese reverse mergers. Working paper.

Choi, Y., Lin, Y., (2009), Consumer response to Mattel product recalls posted on online bulletin boards: Exploring two types of emotion, Public Relations Review, Vol. 35, pp. 18-22.

Chowdhury, A., Frieder, O., Grossman, D., McCabe, M., C., (2002), Collection statistics for fast duplicate document detection. ACM Transactions on Information Systems (TOIS), 20(2):171–191.

Cimiano, P., (2006), Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Secaucus, NJ, USA: Springer-Verlag New York.

Clarkson, M.B.E., (1995), A stakeholder framework for analyzing and evaluating corporate social performance, *Academy of Management Review*, Vol. 20, pp. 92–117.

Clauset, A., Moore, C., Newman, M.E.J., (2004), Finding community structure in very large networks. http://arxiv.org/abs/cond-mat/0408187v2.

Clews, R., Panigirtzoglou, N., Proudman, J., (2000), Recent developments in extracting information from options markets, Bank of England Quarterly Bulletin, February 2000.

Coburn, J., Cook, J., (2014), Cool Response: The SEC & Corporate Climate Change Reporting, Ceres.

Coffee, J. C., (1999), Privatization and corporate governance: The lessons from securities market failure, *Journal of Corporation Law*, Vol. 25, pp. 1-40.

Coffee, J. C., (2002), Racing towards the top: The impact of cross-listing and stock market competition on international corporate governance, *Columbia Law Review,* Vol. 102, pp. 1757-1829.

Coombs, W. T., Holladay, S. J., (1996), Communication and attributions in a crisis: An experimental study of crisis communication, *Journal of Public Relations Research*, Vol. 8, No. 4, pp.279-295.

Coombs, W. T., Holladay, S. J., (2007), The negative communication dynamic: Exploring the impact of stakeholder affect on behavioral intentions, *Journal of Communication Management*, Vol. 11, No. 4, pp. 300-312.

Cooper, C. L., Cartwright, S., Cartright, S., Earley, C. P., (2001), The International Handbook of Organizational Culture and Climate, John Wiley and Sons Ltd.

Coster, E. A., (1992), The perceived quality of working life and job facet satisfaction, *Journal of Industrial Psychology*, Vol. 18, No. 2, pp. 6-9.

Da, Z., Engelberg, J., Gao, P., (2011), In search of attention, *Journal of Finance*, Vol. 66, No. 5, pp. 1461-1499.

Dadvar, M., Hauff, C. De Jong, F. (2011), Scope of negation detection in sentiment analysis. In: Dutch-Belgian Information Retrieval Workshop, DIR 2011.

Daines, R. M., Gow, I., D., Larcker, D. F., (2010), Rating the Ratings: How Good are Commercial Governance Ratings?, *Journal of Financial Economics*, Vol 98, No. 3, pp. 439-461.

Danker, D. J., Luecke, M. M, (2005), Background on FOMC Meeting Minutes, Federal Reserve Bulletin 175-179.

Danthine, J. P., (2013), Causes and consequences of low interest rates, Swisscanto Market Outlook, Lausanne.

Darrough, M., Huang, R., Zhao, S., (2012), The spillover effect of Chinese reverse merger frauds: Chinese or reverse merger? Working paper, Baruch College - City University of New York.

Das, S., Chen, M., (2006), Yahoo! for Amazon: Sentiment extraction from small talk on the web, Working paper, Santa Clara University.

Davis, A. K., Piger, J. M., Sedor, L. M., (2006), Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases, Working paper, Federal Reserve Bank of St. Louis.

Deephouse, D. L., (2000), Media reputation as a strategic resource: An integration of mass communication and resource based theories, *Journal of Management*, Vol. 26, pp. 1091-1112.

Deephouse, D.L., Heugens, P.P.M.A.R., (2009), Linking social issues to organizational impact: The role of infomediaries and the infomediary process, *Journal of Business Ethics*, Vol. 86, No. 4, pp. 541-553.

DellaVigna, S., Pollet, J., (2009), Investor inattention and Friday earnings announcements. *Journal of Finance*, Vol. 64, pp. 709-49.

Demers, E., Vega, C., (2010), Soft information in earnings announcement: News or noise? Working Paper, INSEAD.

Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K., (2013), Microblog-Genre Noise and Impact on Semantic Annotation Accuracy, Proceedings of the 24th ACM Conference on Hypertext and Social Media.

Derwall, J., Koedijk, K., Ter Horst, J., (2011), A tale of values-driven and profit-seeking social investors. Journal of Banking & Finance, Vol. 35, No. 8, pp. 2137-2147.

Diamond, D. and Verrecchia, R. (1991), Disclosure, liquidity and the cost of capital, *Journal of Finance*, Vol. 46, pp. 1325-1360.

Doidge, C., Karolyi, G. A., Stulz, R. M., (2004), Why are foreign firms listed in the U.S. worth more?, *Journal of Financial Economics,* Vol. 71, No. 2, pp. 205-238.

Domowitz, I., Glen, J., Madhavan, A. (1998), International Cross-Listing and Order Flow Migration: Evidence from an Emerging Market, *Journal of Finance*, Vol. 53, No. 6, pp. 2001-2027.

Dougal, C., Engelberg, J., García, D., Parsons, C., (2012), Journalists and the stock market, *Review of Financial Studies,* Vol. 25, No. 4, pp. 639-679.

Dyck, A., Volchkova, N., Zingales, L., (2008), The Corporate Governance Role of the Media: Evidence from Russia, *Journal of Finance,* Vol. 63, pp. 1093-1135.

Easterwood, J., Nutt, S., (1999), Inefficiency in analysts' earnings forecasts: Systematic misreaction or systematic optimism?, *Journal of Finance,* Vol. 54, pp. 1777-1797.

Eberle, D., Berens, G.A.J.M., Li, T., (2013), The impact of interactive corporate social responsibility communication on corporate reputation, *Journal of Business Ethics*, Vol. 118, No. 4, pp. 731-746.

Edmans, A., (2011), Does the stock market fully value intangibles? Employee satisfaction and equity prices, *Journal of Financial Economics, Vol.* 101, pp. 621-640.

Elahi, M., F., Monachesi, P., (2012), An Examination of Cross-Cultural Similarities and Differences from Social Media Data with respect to Language Use, Proceedings of the Eight International Conference on Language Resources and Evaluation.

Emile-Geay, J., Eshleman, J. A., (2013), Toward a semantic web of paleoclimatology. Geochemistry, Geophysics, Geosystems, Vol. 14.

Engelberg, J., (2008), Costly information processing: Evidence from earnings announcements, Working paper, Northwestern University.

Engelberg, J. E., Parsons, C. A., (2011), The causal impact of media in financial markets, *Journal of Finance*, Vol. 66, No. 1, pp. 67-97.

Fama, E., (1965), The Behavior of Stock Market Prices, *Journal of Business*, Vol. 38.

Fama, E. F., French, K. R., (1992), The cross-section of expected stock returns, *Journal of Finance*, Vol. 47, pp. 427-465.

Fama, E. F., French, K. R., (1993), Common risk factors in the returns of stocks and bonds, *Journal of Financial Economics,*Vol. 33, pp. 3-56.

Fang, L. H., Peress, J., (2009), Media coverage and the cross-section of stock returns, *Journal of Finance*, Vol. 64, No. 5, pp. 2023-2052.

Fay, C., Gravelle, T., (2010), Has the Inclusion of Forward-Looking Statements in Monetary Policy Communications Made the Bank of Canada More Transparent?, Bank of Canada Discussion Paper No. 2010-15.

Feng, L., Seasholes, M. S., (2005), Do Investor Sophistication and Trading Experience Eliminate Behavioral Biases in Financial Markets? *Review of Finance,* Vol 9, pp. 305-51.

Ferguson, N. J. (2015), Investor Information Processing and Trading Volume, *Asia-Pacific Journal of Financial Studies*,Vol. 44, No. 2, pp. 322-351.

Ferreira, M. A., Laux, P. A., (2007), Corporate Governance: Idiosyncratic Risk and Information Flow, *Journal of Finance*, Vol. 6, No. 2, pp. 951- 990.

Festinger, L., (1957), A Theory of Cognitive Dissonance, Stanford, CA: Stanford University Press.

Firth, J. R., (1957), Modes of meaning (Essays and Studies) in Firth (1957) 190-215.

Fiske, S. T., Taylor, S. E., (2008), Social cognition: From brains to culture, New York: McGraw-Hill.

Flammer, C., (2013b), Does corporate social responsibility lead to superior financial performance? A regression discontinuity approach, Working paper, University of Western Ontario.

Floyd, K., Voludakis, M., (1999), Attributions for expectancy violating changes in affectionate behavior in platonic friendships, *Journal of Psychology*, Vol. 133, pp. 32-49.

Fombrun, C. J., Shanley, M., (1990), What's in a Name? Reputation Building and Corporate Strategy, *Academy of Management Journal*, Vol. 33, pp. 233-258.

Fombrun, C. J., (1996) Reputation, Harvard Business School Press, Boston.

Fombrun, C. J., (2005), Building corporate reputation through CSR initiatives: Evolving standards, *Corporate Reputation Review*, Vol. 8, No. 1, pp. 7-11.

Fombrun, C., van Riel, C., (1997), The Reputational Landscape, *Corporate Reputation Review*, Vol. 1, pp. 5-13.

Fombrum, C.J., van Riel, C.B.M., (2003), The Reputation Landscape. In J. Balmer and S. Greyser (Eds.), Revealing the Corporation. Perspectives on identity, image, reputation, corporate branding, and corporate level marketing. London: Routledge.

Francis, J., Soffer, L., (1997), The Relative Informativeness of Analysts' Stock Recommendations and Earnings Forecast Revisions, *Journal of Accounting Research*, Vol. 35, Autumn, pp. 193-212.

Frazzini, A., Lamont, O., (2007), The earnings announcement premium and trading volume, NBER Working Paper No. 13090.

Freeman, R. E., (1984), Strategic management: A stakeholder approach. Boston: Pitman.

Freeman, R. E., Harrison, J. S., Wicks, A. C., (2007), Managing for stakeholders: survival, reputation, and success. New Haven, CT: Yale Univ Press.

Fryxell, G. E., Wang, J. (1994), The Fortune corporate reputation index: Reputation for what?, *Journal of Management*, Vol. 20, pp. 1-14.

Gagnon, M. A., Michael, J. H., (2003), Employee strategic alignment at a wood manufacturer: An exploratory analysis using lean manufacturing, *Forest Products Journal*, Vol. 53, 24-29.

Gaines-Ross, L., (2010). Reputation warfare, *Harvard Business Review*, Vol 88. No. 12, 70-76.

Gentzkow, M., Shapiro, J. M., (2006), Media bias and reputation, *Journal of Political Economy*, Vol. 114, pp. 380-316.

Gerhart, B., Rynes, S. L., (2003), Compensation: Theory, evidence, and strategic implications. Thousand Oaks, CA: Sage.

Ghosh, A., Wagner, E. P., (2014), Are U.S. Cross Listed Chinese Companies Tainted by the Poor Audit Quality of Reverse Mergers? The Role of Depository Banks. Working paper.

Gimpel, K., (2006), Modeling topics, *Information Retrieval*, Vol. 5, pp. 1-23.

Givoly, D., Hayn, C., Lourie, B., (2012), Importing accounting quality: The case of foreign reverse mergers, Working paper, Pennsylvania State University and University of California.

Gneezy, U., Meier, S., Rey-Biel, P., (2011), When and Why Incentives (Don't) Work to Modify Behavior, *Journal of Economic Perspectives*, Vol. 25, No. 4, pp. 191-210.

Gray, R., Kouhy, R., Lavers, S., (1995a), Corporate Social and Environmental Reporting: A Review of the Literature and a Longitudinal Study of UK Disclosure, *Accounting, Auditing and Accountability*, Vol. 8, No. 2, 47-77.

Greenberg, J., Knight, G., (2004), Framing sweatshops: Nike, global production, and the American news media, *Communication and Critical/Cultural Studies*,Vol. 1, No. 2, pp. 151-175.

Griffin, J. M., Kelly, P. J., Nardari, F., (2010), Do market efficiency measures yield correct inferences? A comparison of developed and emerging markets*, Review of Financial Studies*, Vol. 23, pp. 3225-3277.

Griffiths, T. L., Steyvers, M., (2004), Finding scientific topics, Proceedings of the National Academy of Sciences, Vol. 101, pp. 5228-5235.

Grimmer, J., (2010), A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases, *Political Analysis*, Vol 18(1):1-35.

Grinblatt, M., Keloharju, M., (2001), How Distance, Language, and Culture Influence Stockholdings and Trades, *Journal of Finance*, Vol 56, No. 3, pp. 1053-1073.

Groseclose, T., Milyo. J., (2005), A Measure of Media Bias, *Quarterly Journal of Economics,* Vol. 120, No. 4, pp. 1191-237.

Gu, F., Lev, B., (2004), The Information Content of Royalty Income, *Accounting Horizons*, Vol. 18, pp. 1-12.

Guiso, L., Sapienza, P., Zingales, L., (2013), The value of corporate culture, *Journal of Financial Economics*.

Gurun, U. G., Butler, A., (2012), Don't Believe the Hype: Local Media Slant, Local Advertising, and Firm Value, Journal of Finance, Vol. 67, No. 2, pp. 561-598.

Healy, P. M., Palepu, K. G., (2001). Information Asymmetry, Corporate Disclosure and the Capital Markets: A review of the Empirical Disclosure Literature, *Journal of Accounting and Economics*, Vol. 31, No. 1, pp. 405-440.

Hendricks, K. B., Singhal, V. R., (2003), The effect of supply chain glitches on shareholder wealth, *Journal of Operations Management*, Vol. 21, No. 5, pp. 501-522.

Hendry, S., Madeley, A., (2010), Text Mining and the Information Content of Bank of Canada Communications, Staff Working Papers 10-31, Bank of Canada.

Hennes, K., Leone, A., Miller, B., (2008), The Importance of Distinguishing Errors from Irregularities in Restatement Research: The Case of Restatements and CEO/CFO Turnover, *The Accounting Review*, Vol. 83, No. 6, pp. 1487-1519.

Hillman, A. J., Keim, G. D., (2001), Shareholder value, stakeholder management, and social issues: What's the bottom line? *Strategic Management Journal,* Vol. 22, pp. 125-139.

Hirshleifer, D., Shumway, T., (2003), Good day sunshine: Stock returns and the weather*, Journal of Finance*, Vol. 58, No. 3, pp. 1009-32.

Hirshleifer, D., Teoh, S. H., (2003), Limited attention, information disclosure, and financial reporting, *Journal of Accounting and Economics*, Vol. 36, pp. 337-386.

Hirshleifer, D., Teoh, S. H., (2011), Limited Investor Attention and Stock Market Misreactions to Accounting Information. *Review of Asset Pricing Studies,* Vol. 1, No. 1, pp. 35-73.

Hofstede, G., (1980), Culture's consequences. Beverly Hills, CA: Sage.

Hogenboom, A., Bal, M., Frasincar, F., Bal, D., (2012), Towards Cross-Language Sentiment Analysis through Universal Star Ratings, KMO pp. 69-79.

Hogenboom, A., Bal, M., Frasincar, F., Bal, D., Kaymak, U., de Jong, F., (2014), Lexicon-based sentiment analysis by mapping conveyed sentiment to intended sentiment, *International Journal of Web Engineering and Technology*, Vol. 9, No. 2, pp. 125-147.

Hong, H., Lim, T., Stein, J. C., (2000), Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies, *Journal of Finance*, Vol. 55, pp. 265-295.

Hong, L., Davison, B. D., (2010), Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics, ISBN 978-1-4503-0217-3.

Hu, M., Liu., B., (2004), Mining and summarizing customer reviews, Tenth ACM International Conference on Knowledge Discovery and Data Mining.

Huang, A.H, Zang, A.Y., Zheng, R., (2014), Evidence on the information content of text in analyst reports*, The Accounting Review*, Vol. 89, No. 6, pp. 2151-2180.

Huberman, G., Regev, T., (2001), Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar, *Journal of Finance*, Vol. 56, No. 1, pp. 387-396.

Iliev, P., Lins, K., Miller, D., Roth, L., (2015), Shareholder Voting and Corporate Governance Around the World, *Review of Financial Studies*, Vol. 28, No. 8, pp. 2167-2202.

Ivkovic, Z., Weisbenner, S. J., (2005), Local Does as Local Is: Information Content of the Geography of ′ Individual Investors Common Stock Investments, *Journal of Finance*, Vol. 60, pp. 267-306.

Iyengar, S., Kinder, D., (1987), News that mattes: Television and American opinion, Chicago, IL: University of Chicago Press.

Jagarlamudi, J., Daume III, H., Udupa, R., (2012), Incorporating Lexical Priors into Topic Models, EACL.

Jegadeesh, N., Titman, S., (1993), Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance,* Vol. 48, pp. 65-91.

Jegadeesh, N., Wu, D., (2013), Word Power: A New Approach for Content Analysis, *Journal of Financial Economics*, Vol. 110, No. 1, pp. 712-729.

Jensen, M. C., (1986), Agency Costs of Free Cash Flow, Corporate Finance and Takeovers, *American Economic Review*, Vol. 76, No. 2, pp. 323-329.

Jensen, M. C., (2002), Value maximization, stakeholder theory, and the corporate objective function, *Business Ethics Quarterly*, Vol. 12, No. 2, pp. 235-256.

Jensen, M., Meckling, W., (1976), Theory of the firm: managerial behavior, agency costs and ownership structure, *Journal of Financial Economics*, Vol. 3, pp. 305-360.

Jin, Y., Liu, B., F., Austin, L., (2011), Examining the Role of Social Media in Effective Crisis Management: The Effects of Crisis Origin, Information Form, and Source on Publics' Crisis Responses, Communication Research.

Kamenica, E., (2012), Behavioral Economics and Psychology of Incentives, *Annual Review of Economics*, Vol. 4, pp. 427-452.

Kampgen, B., Weller, T., O’Riain, S., Weber , C., Harth, A., (2014), Accepting the XBRL Challenge with Linked Data for Financial Data Integration, ESWC.

Kan, X., Ren, F., (2011), Sampling Latent Emotions and Topics in a Hierarchical Bayesian Network. IEEE NLP-KE.

Kaplan, S. N., Zingales, L., (1997), Do Investment-Cash Flow Sensitivities Provide Useful Measures of Financing Constraints? *Quarterly Journal of Economics*, Vol. 112, pp. 169-215.

Kennedy, A., Inkpen, D., (2006), Sentiment Classification of Movie Reviews using Contextual Valence Shifters, *Computational Intelligence*, Vol. 22, No. 2, pp. 110-125.

Kim, H., Cameron, G., (2011), Emotions Matter in Crisis: The Role of Anger and Sadness in the Publics’ Response to Crisis News Framing and Corporate Crisis Response, Communication Research.

Kim, O., Verrecchia, R. E., (1991), Trading Volume and Price Reactions to Public Announcements, *Journal of Accounting Research*, Vol. 29, No. 2, pp. 302-321.

Kim, S. M., Hovy, E., (2004), Determining the Sentiment of Opinions, Proceedings of the 20th International Conference on Computational Linguistics.

Kilbanoff, P., Lamont, O., Wizman, T., (1998), Investor reaction to salient news in closed-end country funds, *Journal of Finance*, Vol. 53, No. 2, pp. 673-699.

Klein, J., Dawar, N., (2004), Corporate social responsibility and consumers' attributions and brand evaluations in a product-harm crisis, *International Journal of Research in Marketing*, Vol. 21, No. 3, pp. 203-217.

Kolk, A. (2008), Sustainability, accountability and corporate governance: Exploring multinationals reporting practices, *Business Strategy and the Environment*, Vol. 17, No. 1, pp. 1-15.

Kozareva, Z., Navarro, B., Vazquez, S., Nibtoyo, A., (2007), UA-ZBSA: A Headline Emotion Classification through Web Information, Proceedings of the 4th International Workshop on Semantic Evaluations.

Lange, D. A., Washburn, N. T., (2012), Understanding attributions of corporate social irresponsibility, *Academy of Management Review*, Vol. 37, pp. 300-326.

Latham, G. P., Locke, E. A. (2006). Enhancing the benefits and overcoming the pitfalls of goal setting. *Organizational Dynamics*, Vol. 35, No. 4, pp. 332-340.

Lau, J.H., Grieser, K., Newman, D., Baldwin, T., (2011), Automatic labelling of topic models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1536-45.

Lazarus, R. S., (1991), Emotion and adaptation, New York: Oxford University Press.

Lee, S.Y. M, Chen, Y and Huang, C.R., (2010a), A Text driven Rule-based System for Emotion Cause Detection. NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.

Lev, B., (2001), Intangibles: Management, Measurement, and Reporting, Brookings Institution Press.

Lev, B., Sougiannis, T., (1996), The capitalization, amortization, and value relevance of R&D, *Journal of Accounting and Economics,* Vol. 21, pp. 107-138.

Levering, R., Moskowitz, M., Katz, M., (1984), The 100 Best Companies to Work for in America. Addison-Wesley, Reading, MA.

Lev, B., Radhakrishnan, S., Zhang, W., (2009), Organization capital, Abacus, pp. 275-298.

Li, F., (2006), Do stock market investors understand the risk sentiment of corporate annual reports? Working paper, University of Michigan.

Li, T., Berens, G., de Maertelaere, M., (2013), Corporate Twitter Channels: The Impact of Engagement and Informedness on Corporate Reputation, *International Journal of Electronic Commerce*, Vol. 18, No. 2, pp. 97-126.

Licht, A., (2003), Cross-listing and corporate governance: Bonding or avoiding?, *Chicago Journal of International Law*, Vol. 4, pp. 141-163.

Litvak, M., Last, M., (2008), Graph-based keyword extraction for singledocument summarization, Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, pp. 17-24.

Liu, B., (2012), Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies. Morgan & Claypool, Vol. 16.

Liu, Y., Niculescu-Mizil, A., Gryc, W., (2009), Topic-link LDA: joint models of topic and author community, Proceedings of the 26th Annual International Conference on Machine Learning, pp. 665-672.

Lobel, O., (2007), Big-box benefits: The targeting of giants in a national campaign to raise work conditions, *Connecticut Law Review*, Vol. 39, pp. 1685-1712.

Locke, E.A., Latham, G.P, (1984), Goal setting: A motivational technique that works, Englewood Cliffs N.J: Prentice- Hall.

Locke, E. A., Latham, G.P., (2002), Building a practically useful theory of goal setting and task Motivation: A 35-year odyssey, *American Psychology*, Vol. 57, pp. 705-717.

Locke, E. A., (1966), The relationship of intentions to level of performance, *Journal of Applied Psychology,* Vol. 50, pp. 60-66.

Locke, E. A., Latham, G. P., (1990), A theory of goal setting and task performance, Englewood Cliffs, NJ: Prentice Hall.

Locke, E. A., Latham, G. P., (2006), New directions in goal-setting theory*, Association for Psychlogical Science*, Vol. 15, No. 5, pp. 265-270.

Locke, E. A., Latham, G. P., (2007), New developments in and directions for goal-setting research, *European Psychologist*, Vol. 12, No. 4, pp. 290-300.

Loughran, T., McDonald, B., (2011), When is a liability not a liability? Textual analysis, dictionaries and 10Ks. *Journal of Finance*, Vol. 66, pp. 35-65.

MacKinlay, A. C., (1997), Event Studies in Economics and Finance, *Journal of Economic Literature*, Vol. 35, No. 1, pp. 13-39.

Madden, T.J., Fehle, F., Fournier, S.M., (2006), Brands Matter: An Empirical Investigation of Brand Building Activities and the Creation of Shareholder Value, *Journal of the Academy of Marketing Science*, Vol. 34, No. 2, pp. 224-235.

Mahon, J. F., (2002), Corporate Reputation: A Research Agenda Using Strategy and Stakeholder Literature, *Business & Society*, Vol. 41, No. 4, pp. 415- 445.

Mangold, W., Faulds, D., (2009), Social media: The new hybrid element of the promotion mix, *Business Horizons*, Vol. 52, No. 4, pp. 357-365.

Manning, C., Schütze, H., (1999), Foundations of statistical natural language processing MIT Press, Cambridge, MA.

Manning, C.D, Raghavan, P., Schütze, H., (2008), Introduction to Information Retrieval, Cambridge University Press.

Marcus, A. A., Goodman, R. S., (1991), Victims and shareholders: The dilemmas of presenting corporate policy during a crisis, *Academy of Management Journal*, Vol. 34.

Maritz, J. S., (1981), Distribution-Free Statistical Methods. London: Chapman & Hall.

Marquis, C., Toffel, M., (2012), When Do Firms Greenwash? Corporate Visibility, Civil Society Scrutiny, and Environmental Disclosure, Harvard Business School Discussion Paper 12-43.

Mayew, W.J., Venkatachalam, M., (2011), The power of voice: Managerial affective states and future firm performance, *Journal of Finance*, Vol. 67, No. 1, pp. 1-44.

McCallum, A., Nigam, K., (1998), A comparison of event models for Naive Bayes text classification. Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, 41-48.

McCombs, M., Shaw, D., (1972), The agenda-setting function of mass media, *Public Opinion Quarterly*, No. 36, pp. 963-975.

McDonnell, M.H., King, B., (2013), Keeping up Appearances Reputational Threat and Impression Management after Social Movement Boycotts, *Administrative Science Quarterly*, Vol. 58, No. 3, pp. 387-419.

Melillo, J. M., Richmond, T.T., Yohe, G.W., (2014), Climate Change Impacts in the United States: The Third National Climate Assessment, U.S. Global Change Research Program.

Meyersson, P., Karlberg, P.P., (2012), A Journey in Communication: the Case of the Sveriges Riksbank, SNS Förlag.

Mihalcea, R., and Tarau, P., (2004), TextRank: Bringing order into texts, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.

Miller, G. S., (2006), The press as a watchdog for accounting fraud, *Journal of Accounting Research*, Vol. 44, No. 5, pp.1001-1033.

Mimno, D., (2012), Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage*, Vol. 5, No(1):3.

Moniz, A., de Jong, F., (2014a), Sentiment Analysis and the Impact of Employee Satisfaction on Firm Earnings. Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Springer 2014 Lecture Notes in Computer Science.

Moniz, A., de Jong, F., (2014b), Predicting the impact of central bank communications on financial market investors' interest rate expectations, The Semantic Web: ESWC 2014. Springer 2014 Lecture Notes in Computer Science.

Moniz, A., de Jong, F., (2014c), Classifying the impact of negative affect expressed by the financial media on investor behavior, Proceedings of the 6th Conference on Information Interaction in Context (IIiX).

Moniz, A., de Jong, F., (2014d), Reputational DAMAGE: Classifying the impact of allegations of irresponsible corporate behavior expressed in the financial media, 34th International Symposium on Forecasting 2014 conference proceedings.

Moniz, A., de Jong, F., (2015), Analysis of companies' non-financial disclosures: Ontology learning by topic modeling. The Semantic Web: ESWC 2015, Springer 2015 Lecture Notes in Computer Science.

Mullainathan, S., Shleifer, A., (2005), The market for news, *American Economic Review*, Vol. 95, No. 4, pp. 1031-1053.

Murphy, D., Shrieves, R., Tibbs, S., (2009), Understanding the Penalties Associated with Corporate Misconduct: An Empirical Examination of Earnings and Risk*, Journal of Financial and Quantitative Analysis*, Vol. 43, pp.581-612.

Newman, D., Lau, J.H., Grieser, K., Baldwin, T., (2010), Automatic Evaluation of Topic Coherence, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 100-108.

Noble, C. H., (1999), The eclectic roots of strategy implementation research, *Journal of Business Research*, Vol. 45, pp. 119-34.

Opsahl, T., Agneessens, F., Skvoretz, J., (2010), Node centrality in weighted networks: Generalizing degree and shortest paths, *Social Networks*, Vol. 32, No. 3, pp. 245-251.

O'Riain, S., Curry, E., Harth, A., (2012), XBRL and open data for global financial ecosystems: A linked data approach, International Journal of Accounting Information Systems.

O'Reilly, C., Chatman, J., Caldwell, D., (1991), People and Organizational Culture: A Profile Comparison Approach to Assessing Person-Organization Fit, *Academy of Management Journal*, Vol. 34, pp. 487-516.

Ordóñez, L. D., Schweitzer, M. E., Galinsky, A. D., Bazerman, M. H., (2009), On Good Scholarship, Goal Setting, and Scholars Gone Wild, *Academy of Management Perspectives*, Vol. 23, No. 1, pp. 6-16.

Orlitzky, M., Benjamin, J.D. (2001). Corporate Social Performance and Firm Risk: A Meta-Analytic Review. Business and Society, Vol. 40, No. 4, pp. 369-396.

Ozik, G., Sadka, R., (2013), Media coverage and hedge-fund returns, *Financial Analysts Journal*, Vol. 69, May/June 2013, pp. 57-75.

Pang, B., Lee, L., and Vaithyanathan, S., (2002), Thumbs up? Sentiment classification using machine learning techniques, Proceedings of EMNLP-02.

Pang, B., Lee, L., (2004), A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts, 42nd Annual Meeting of the Association for Computational Linguistics, pp. 271–280.

Peetz, M. H., (2015), Time-aware online reputation analysis, University of Amsterdam.
http://dare.uva.nl/record/1/468533

Pennebaker, J. W., Mehl, M. R., Niederhoffer, K. G., (2003), Psychological aspects of natural language. use: our words, our selves, *Annual review of psychology*, Vol. 54, No. 1, pp. 547-577.

Peress, J, (2014), The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes, *Journal of Finance*, Vol. 69, No. 5, pp. 2007-2043.

Petersen, M., (2004), Information: Hard and Soft, Working Paper.

Petersen, M., (2009), Estimating standard errors in finance panel data sets: comparing approaches, *Review of Financial Studies*, Vol 22, pp. 435-480.

Pfarrer, M., Pollock, T., Rindova, V., (2010), A tale of two assaets: The effects of firm reputational and celebrity on earnings surprises and investors' reactions.

Pinder, C. C., (1998), Work motivation in organizational behavior, Upper Saddle River, NJ: Prentice-Hall.

Plutchik, R., (1962), The emotions: Facts, theories, and a new model, New York: Random House.

Plutchik, R., (1980), A general psychoevolutionary theory of emotion, R. Plutchik & H. Kellerman (Eds.), Emotion: Theory, research, and experience, Vol. 1. Theories of emotion.

Podesta, J., Pritzker, P., Moniz, E., Holdern, J., Zients, J., (2014), Big data: Seizing opportunities, preserving values, Washington, DC: Executive Office of the President. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Polanyi, L., Zaenen, A., (2004), Contextual Valence Shifters, AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications.

Pollock, T. G., Rindova, V. P., (2003), Media legitimation effects in the market for initial public offerings, *Academy of Management Journal*, Vol. 46, pp. 631-642.

Popadak, J., (2013), A Corporate Culture Channel: How Increased Shareholder Governance Reduces Firm Value, Duke University working paper.

Pouchard, L., Branstetter, M., Cook, R., Devarakonda, R., Green, J., Palanisamy, G., (2013), A Linked Science Investigation: Enhancing Climate Change Data Discovery with Semantic Technologies, Oak Ridge National Laboratory.

Price, S. M., Doran, J. S., Petersen, D. R., Bliss, B. A., (2012), Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone, *Journal of Banking and Finance*, Vol. 36, No. 4, pp. 992-1011.

Qin, Z., Thint, M., Huang, Z., (2009), Ranking Answers by Hierarchical Topic Models, 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2009.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., McFarland, D. A., (2009), Topic modeling for the social sciences, NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond.

Ramage, D., Dumais, S., Liebling, D., (2010) Characterizing microblogs with topic models, ICWSM.

Rao, H., Yue, L. Q., Ingram, P., (2011), Laws of Attraction. American Sociological Review, 76(3): 365-385.

Rhee, M., Haunschild, P., (2006), The liability of good reputation: A study of product recalls in the U.S. automobile industry, *Organization Science*, Vol. 17, No. 1, pp. 101-117.

Russo, M.V., Fouts, P.A., (1997), A resource-based perspective on corporate environmental performance and profitability*, Academy of Management Journal*, Vol. 40, No. 3, pp. 534–559.

Schouten, K., Frasincar, F., (2015), Survey on Aspect-Level Sentiment Analysis, IEEE Transactions on Knowledge and Data Engineering, Vol. 28 , No. 3, pp 813-830

Schweitzer, M. E., Ordóñez, L., Douma, B., (2004), Goal Setting as a Motivator of Unethical Behavior, *Academy of Management Journal*, Vol. 47, No. 3, pp. 422-432.

Shane, P., Spicer, B. (1983). Market response to environmental information produced outside the firm, *Accounting Review*, Vol. 58, No. 3, pp. 521–538.

Shleifer, A., (1986), Do demand curves for stocks slope down?, *Journal of Finance*, Vol. 41, pp. 579–590.

Siegel, J., (2005), Can Foreign Firms Bond themselves Effectively by Renting US Securities Laws? *Journal of Financial Economics*, Vol. 75, No. 2, pp. 319-359.

Skinner, D.J., (2008a), Accounting for Intangibles - a critical review of policy recommendations, Accounting and Business Research, Vol. 38, No. 3, pp. 191-204.

Solomon, D., (2012), Selective publicity and stock prices, *Journal of Finance*, Vol. 67, pp. 599-637.

Spector, P.E., (2003), Industrial and organizational psychology- Research and practice (3rd ed.), New York: John Wiley & Sons, Inc.

Statman, M., Glushkov, D., (2009), The Wages of Social Responsibility, *Financial Analysts Journal*, Vol. 65, pp. 33–46.

Stone, P., Dumphy, D. C., Smith, M. S., Ogilvie, D. M., (1966), The General Inquirer: A Computer Approach to Content Analysis, The MIT Press.

Strapparava, C., Mihalcea, R., (2008), Learning to identify emotions in text, Proceedings of the 2008 ACM symposium on Applied Computing, pp. 1556-1560.

Stulz, R. M., (1999), Globalization, corporate finance, and the cost of capital, *Journal of Applied Corporate Finance*, Vol. 12, No. 3, pp. 8-25.

Suchman, M. C. 1995. Managing legitimacy: Strategic and institutional approaches, *Academy of Management Review*, Vol. 20, pp. 571-610.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., Zhang, M., (2014), Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis, Proceedings of the 31st International Conference on Machine Learning.

Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M., (2006), Hierarchical Dirichlet processes, *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581.

Tetlock, P. C., (2007), Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance*, Vol. 62, pp. 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S., (2008), More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance*, Vol. 63, pp. 1437-1467.

Tilly, C., (2007), Wal-Mart and its workers: Not the same all over the world, *Connecticut Law Review*, Vol. 39, pp. 1805-1823.

Tilmes, C., Fox, P., Ma, X., McGuinness, D. L., Privette, A.P., Smith, A.,Waple, A., Zednik, S., Zheng, J. G., (2013), Provenance Representation for the National Climate Assessment in the Global Change Information System, IEEE.

Titov, I, McDonald, R.T., (2008), Modeling online reviews with multi-grain topic models, Proceedings of the 17th international conference on World Wide Web, 111-120, 253.

Triandis, H. C, Bontempo, R., Villareal, M. J., Asai, M., Lucca, N., (1988), Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships*, Journal of Personality and Social Psychology*, Vol. 54, pp. 323-338.

Tziner, A., Latham, G. P., (1989), The effects of appraisal instrument, feedback and goal-setting on worker satisfaction and commitment, *Journal of Organizational Behavior*, Vol. 10, pp. 145-153.

van Riel, C., (1995), Principles of Corporate Communication, Prentice Hall PTR.

Vanhamme, J., Grobben, B., (2008), Too Good to be True! The Effectiveness of CSR History in Countering Negative Publicity, *Journal of Business Ethics*, Vol. 85, pp. 273-283.

Vayid, I., (2013), Central Bank Communications Before, During and After the Crisis: From Open-Market Operations to Open-Moth Policy, Bank of Canada Working Paper 2013-41.

Verwijmeren, P., Derwall, J., (2010), Employee Well-Being, Firm Leverage, and Bankruptcy Risk, *Journal of Banking and Finance*, Vol. 34, No. 5, pp. 956-964.

Viñals, J. Lessons from the Crisis for Central Banks. IMF Speech (2010).

Waddock, S. A., S. B. Graves, (1997), The corporate social performance- financial performance link, *Strategic Management Journal*, Vol. 18, No. 4, pp. 303-319.

Wagner, C. M., (2007), Organizational commitment as a predictor variable in nursing turnover research: Literature review, *Journal of Advanced Nursing*, Vol. 60, No. 3, pp. 235-247.

Wallach, H, Mimno, D., McCallum, A., (2009), Rethinking LDA: Why Priors Matter, NIPS, 2009.

Wang, X., McCallum A., Wei, X., (2007), Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval, Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), 2007.

Wartick, S., L., (1992), The relationship between intense media exposure and change in corporate reputation, *Business & Society*, Vol. 31, No. 1, pp. 33-49.

Watson, D., Clark, L.A., (1984), Negative affectivity: The disposition to experience negative aversive emotional states, *Psychological Bulletin*, Vol. 96, pp. 465–490.

Weaver, D. H., (1980), Audience need for orientation and media effects. Communication Research

Wei, W., Barnaghi, P., Bargiela, A., (2009), Probabilistic Topic Models for Learning Terminological Ontologies, IEEE.

Weiner, B., (1985), An attribution theory of achievement motivation and emotion, *Psychological Review*, Vol. 97, pp. 548-573.

Wierzbicka, A., (1995), Emotion and Facial Expression: A Semantic Perspective, *Culture Psychology*, Vol. 1, No. 2, pp. 227–258.

Wiese, B. S., Freund, A. M., (2005), Goal progress makes one happy, or does it? Longitudinal findings from the work domain, *Journal of Occupational and Organizational Psychology*, Vol. 78, pp. 287–304.

Wiesenfeld, B. M., Wurthmann, K. A., Hambrick, D. C., (2008), The stigmatization and devaluation of elites associated with corporate failure: A process model, *Academy of Management Review*, Vol. 33, No.1, pp. 231–251.

Witt, L. A., (1998), Enhancing Organizational Goal Congruence: A solution to Organizational Politics, *Journal of Applied Psychology*, Vol. 83, pp. 666-74.

Wong, W., Liu, W., Bennamoun, M., (2011), Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances, IGI Global, USA.

Yukl, G. A., Latham, G. P., (1978), Interrelationships among employee participation, individuals differences, goal difficulty, goal acceptance, goal instrumentality, and performance, *Personnel Psychology*, Vol. 31, pp. 305-323.

Zavitsanos, E., Paliouras, G., Vouros, G.A., Petridis, S., (2007), Discovering subsumption hierarchies of ontology concepts from text corpora, Proceedings of the International Conference on Web Intelligence.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E., Yan, H Li, X, (2011), Comparing Twitter and traditional media using topic models, *Advances in Information Retrieval*, pp.338-349.

Zhu, Q., (2009), The Home Stigma: Adverse Selection in ADRs and the Home Capital Market Environment, Proceedings of 2009 FMA Annual Meeting, 1-40.

Zipf, G. K., (1932), Selected Studies of the Principle of Relative Frequency in Language, Cambridge, MA.: Harvard University Press.

Summary

Traditionally, equity investors have relied upon the information reported in firms' financial accounts to make their investment decisions. Due to the conservative nature of accounting standards, firms cannot value their intangible assets such as corporate culture, brand value and reputation. Investors' efforts to collect such information have been hampered by the voluntary nature of Corporate Social Responsibility (CSR) reporting standards, which have resulted in the publication of inconsistent, stale and incomplete information across firms. In short, information on intangible assets is less salient to investors compared to accounting information because it is more costly to collect, process and analyse.

In this thesis we design an automated approach to collect and quantify information on firms' intangible assets by drawing upon techniques commonly adopted in the fields of Natural Language Processing (NLP) and Information Retrieval. The exploitation of unstructured data available on the Web holds promise for investors seeking to integrate a wider variety of information into their investment processes. The objectives of this research are: 1) to draw upon textual analysis methodologies to measure intangible information from a range of unstructured data sources, 2) to integrate intangible information and accounting information into an investment analysis framework, 3) evaluate the merits of unstructured data for the prediction of firms' future earnings.

Nederlandse Samenvatting (Summary in Dutch)

Traditioneel gezien zijn aandelenbeleggers bij het nemen van hun investeringsbeslissingen uitgegaan van de informatie die gerapporteerd wordt in de financiële rekeningen van bedrijven. Als gevolg van de conservatieve aard van boekhoudkundige normen kunnen bedrijven echter hun immateriële activa - zoals bedrijfscultuur, merkwaarde en reputatie - niet waarderen in deze rekeningen. Inspanningen van beleggers om dergelijke informatie te verzamelen worden gehinderd door het vrijwillige karakter van standaarden voor de rapportage van Maatschappelijk Verantwoord Ondernemen (MVO), die hebben geleid tot de wijdverbreide publicatie van inconsistente, verouderde en onvolledige informatie. Kortweg is informatie over immateriële activa minder saillant voor beleggers dan boekhoudkundige informatie, omdat het verzamelen, verwerken en analyseren ervan hogere kosten met zich meebrengt.

In dit proefschrift ontwerpen we een geautomatiseerde aanpak voor het verzamelen en kwantificeren van informatie over de immateriële activa van bedrijven, door gebruik te maken van technieken die gewoonlijk op het gebied van Natural Language Processing (NLP) en Information Retrieval worden toegepast. De exploitatie van ongestructureerde gegevens beschikbaar op het Web is veelbelovend voor beleggers die ernaar streven een grotere verscheidenheid aan informatie te integreren in hun beleggingsprocessen. De doelstellingen van dit onderzoek zijn: 1) gebruik te maken van tekstuele analysemethoden om immateriële informatie te meten uit een scala aan ongestructureerde gegevensbronnen, 2) immateriële informatie en boekhoudkundige informatie te integreren in een raamwerk voor beleggingsanalyse, 3) de verdiensten van ongestructureerde data te evalueren voor het voorspellen van toekomstige winsten van bedrijven.

About the Author

Andy Moniz works at UBS and is responsible for the design of systematic equity strategies using unstructured data. Andy studied at Downing College, University of Cambridge, between 1997-2000. He graduated with a BA and MA in Economics and received a university scholarship and college prize. Andy began his career in 2000 as a macroeconomist at the Bank of England where he worked within the Conjunctural Assessment and Projections Division on the design of the Bank's macroeconomic forecasting model. Between 2003-2011, Andy worked within the Quantitative Equity Strategies departments for various investment banks. During this time he became a CFA Charterholder and undertook a part-time MSc in Statistics from the University of London. In 2011, Andy moved to the Netherlands to work as a senior quantitative portfolio manager at APG Asset Management and was responsible for the design of stock selection strategies. His academic research has been presented at numerous conferences in the fields of computational linguistics, information retrieval and finance.

Portfolio

Publications

Moniz, A., de Jong, F., (2014a), Sentiment Analysis and the Impact of Employee Satisfaction on Firm Earnings. Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Springer 2014 Lecture Notes in Computer Science.

Moniz, A., de Jong, F., (2014b), Predicting the impact of central bank communications on financial market investors' interest rate expectations, The Semantic Web: ESWC 2014. Springer 2014 Lecture Notes in Computer Science.

Moniz, A., de Jong, F., (2014c), Classifying the impact of negative affect expressed by the financial media on investor behavior, Proceedings of the 6th Conference on Information Interaction in Context (IIiX).

Moniz, A., de Jong, F., (2014d), Reputational DAMAGE: Classifying the impact of allegations of irresponsible corporate behavior expressed in the financial media, 34th International Symposium on Forecasting 2014 conference proceedings.

Moniz, A., de Jong, F., (2015), Analysis of companies' non-financial disclosures: Ontology learning by topic modeling. The Semantic Web: ESWC 2015, Springer 2015 Lecture Notes in Computer Science.

INSTITUTE OF MANAGEMENT (ERIM)

ERIM PH.D. SERIES RESEARCH IN MANAGEMENT

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl/pub. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

DISSERTATIONS LAST FIVE YEARS

Abbink, E.J., *Crew Management in Passenger Rail Transport*, Promotor(s): Prof.dr. L.G. Kroon & Prof.dr. A.P.M. Wagelmans, EPS-2014-325-LIS, http://repub.eur.nl/ pub/76927

Acar, O.A., *Crowdsourcing for Innovation: Unpacking Motivational, Knowledge and Relational Mechanisms of Innovative Behavior in Crowdsourcing Platforms*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2014-321-LIS, http://repub.eur.nl/pub/76076

Akin Ates, M., *Purchasing and Supply Management at the Purchase Category Level: strategy, structure and performance*, Promotor(s): Prof.dr. J.Y.F. Wynstra & Dr. E.M. van Raaij, EPS-2014-300-LIS, http://repub.eur.nl/pub/50283

Akpinar, E., *Consumer Information Sharing*, Promotor(s): Prof.dr.ir. A. Smidts, EPS- 2013-297-MKT, http://repub.eur.nl/pub/50140

Alexander, L., *People, Politics, and Innovation: A Process Perspective*, Promotor(s): Prof.dr. H.G. Barkema & Prof.dr. D.L. van Knippenberg, EPS-2014-331-S&E, http: //repub.eur.nl/pub/77209

Almeida e Santos Nogueira, R.J. de, *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*, Promotor(s): Prof.dr.ir. U. Kaymak & Prof.dr. J.M.C. Sousa, EPS-2014-310-LIS, http://repub.eur.nl/pub/51560

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-frequency Data*, Promotor(s): Prof.dr. D.J.C. van Dijk, EPS-2013-273-F&A, http://repub.eur.nl/pub/38240

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promotor(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2013-278-S&E, http://repub.eur.nl/pub/39128

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, http://repub.eur.nl/pub/23670

Benschop, N, *Biases in Project Escalation: Names, frames & construal levels*, Promotors: Prof.dr. K.I.M. Rhode, Prof.dr. H.R. Commandeur, Prof.dr. M.Keil & Dr. A.L.P. Nuijten, EPS-2015-375-S&E, hdl.handle.net/1765/79408

Berg, W.E. van den, *Understanding Salesforce Behavior using Genetic Association Studies*, Promotor(s): Prof.dr. W.J.M.I. Verbeke, EPS-2014-311-MKT, http://repub.eur.nl/pub/51440

Betancourt, N.E., *Typical Atypicality: Formal and Informal Institutional Conformity, Deviance, and Dynamics*, Promotor(s): Prof.dr. B. Krug, EPS-2012-262-ORG, http://repub.eur.nl/pub/32345

Bliek, R. de, *Empirical Studies on the Economic Impact of Trust*, Promotor(s): Prof.dr. J. Veenman & Prof.dr. Ph.H.B.F. Franses, EPS-2015-324-ORG, http://repub.eur.nl/pub/78159

Blitz, D.C., *Benchmarking Benchmarks*, Promotor(s): Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, http://repub.eur.nl/pub/22624

Boons, M., *Working Together Alone in the Online Crowd: The Effects of Social Motivationsand Individual Knowledge Backgrounds on the Participation and Performance of Members of Online Crowdsourcing Platforms*, Promotor(s): Prof.dr. H.G. Barkema & Dr. D.A. Stam, EPS-2014-306-S&E, http://repub.eur.nl/pub/50711

Brazys, J., *Aggregated Marcoeconomic News and Price Discovery*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2015-351-F&A, http://repub.eur.nl/pub/78243

Burger, M.J., *Structure and Cooptition in Urban Networks*, Promotor(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.R. Commandeur, EPS-2011-243-ORG, http://repub.eur.nl/pub/26178

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit,Coworker Satisfaction, and Relational Models*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-292-ORG, http://repub.eur.nl/pub/41508

Camacho, N.M., *Health and Marketing: Essays on Physician and Patient Decision- Making*, Promotor(s): Prof.dr. S. Stremersch, EPS-2011-237-MKT, http://repub.eur.nl/pub/23604

Cancurtaran, P., *Essays on Accelerated Product Development*, Promotor(s): Prof.dr. F. Langerak & Prof.dr.ir. G.H. van Bruggen, EPS-2014-317-MKT, http://repub.eur.nl/pub/76074

Caron, E.A.M., *Explanation of Exceptional Values in Multi-dimensional Business Databases*, Promotor(s): Prof.dr.ir. H.A.M. Daniels & Prof.dr. G.W.J. Hendrikse, EPS-2013-296-LIS, http://repub.eur.nl/pub/50005

Carvalho, L. de, *Knowledge Locations in Cities: Emergence and Development Dynamics*, Promotor(s): Prof.dr. L. Berg, EPS-2013-274-S&E, http://repub.eur.nl/pub/38449

Consiglio, I., *Others: Essays on Interpersonal and Consumer Behavior*, Promotor: Prof.dr. S.M.J. van Osselaer, EPS-2016-366-MKT, http://repub.eur.nl/pub/79820

Cox, R.H.G.M., *To Own, To Finance, and To Insure - Residential Real Estate Revealed*, Promotor(s): Prof.dr. D. Brounen, EPS-2013-290-F&A, http://repub.eur.nl/pub/40964

Deichmann, D., *Idea Management: Perspectives from Leadership, Learning, and Network Theory*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2012-255-ORG, http://repub.eur.nl/pub/31174

Deng, W., *Social Capital and Diversification of Cooperatives*, Promotor(s): Prof.dr. G.W.J. Hendrikse, EPS-2015-341-ORG, http://repub.eur.nl/pub/77449

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promotor(s): Prof.dr. D. de Cremer, EPS-2011-232-ORG, http://repub.eur.nl/pub/23268

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promotor(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-272-LIS, http://repub.eur.nl/pub/38241

Doorn, S. van, *Managing Entrepreneurial Orientation*, Promotor(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch, & Prof.dr. H.W. Volberda, EPS-2012-258-STR, http://repub.eur.nl/pub/32166

Douwens-Zonneveld, M.G., *Animal Spirits and Extreme Confidence: No Guts, No Glory?*Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2012-257-F&A, http://repub.eur.nl/pub/31914

Duca, E., *The Impact of Investor Demand on Security Offerings*, Promotor(s): Prof.dr. A. de Jong, EPS-2011-240-F&A, http://repub.eur.nl/pub/26041

Duyvesteyn, J.G. *Empirical Studies on Sovereign Fixed Income Markets,* Promotor(s): Prof.dr P.Verwijmeren & Prof.dr. M.P.E. Martens, EPS-2015-361-F&A, hdl.handle.net/1765/79033

Duursema, H., *Strategic Leadership: Moving Beyond the Leader-Follower Dyad*, Promotor(s): Prof.dr. R.J.M. van Tulder, EPS-2013-279-ORG, http://repub.eur.nl/pub/39129

Eck, N.J. van, *Methodological Advances in Bibliometric Mapping of Science*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2011-247-LIS, http://repub.eur.nl/pub/26509

Elemes, A, *Studies on Determinants and Consequences of Financial Reporting Quality,* Promotor: Prof.dr. E.Peek, EPS-2015-354-F&A, http://hdl.handle.net/1765/79037

Ellen, S. ter, *Measurement, Dynamics, and Implications of Heterogeneous Beliefs in Financial Markets*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2015-343-F&A, http://repub.eur.nl/pub/78191

Erlemann, C., *Gender and Leadership Aspiration: The Impact of the Organizational Environment*, Promotor: Prof.dr. D.L. van Knippenberg, EPS-2016-376-ORG, http://repub.eur.nl/pub/79409

Eskenazi, P.I., *The Accountable Animal*, Promotor(s): Prof.dr. F.G.H. Hartmann, EPS-2015-355-F&A, http://repub.eur.nl/pub/78300

Essen, M. van, *An Institution-Based View of Ownership*, Promotor(s): Prof.dr. J. Van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, http://repub.eur.nl/pub/22643

Evangelidis, I., *Preference Construction under Prominence*, Promotor(s): Prof.dr. S.M.J. van Osselaer, EPS-2015-340-MKT, http://repub.eur.nl/pub/78202

Faber, N., *Structuring Warehouse Management*, Promotor(s): Prof.dr. MB.M. de Koster, Prof.dr. Ale Smidts, EPS-2015-336-LIS, http://repub.eur.nl/pub/78603

Fernald, K., *The Waves of Biotechnological Innovation in Medicine: Interfirm Cooperation Effects and a Venture Capital Perspective,* Promotor(s): Prof.dr. E.Claassen, Prof.dr. H.P.G.Pennings & Prof.dr. H.R. Commandeur, EPS-2015-371-S&E, http://hdl.handle.net/1765/79120

Fourne, S.P., *Managing Organizational Tensions: A Multi-Level Perspective on Exploration, Exploitation and Ambidexterity*, Promotor(s): Prof.dr. J.J.P. Jansen & Prof.dr. S.J. Magala, EPS-2014-318-S&E, http://repub.eur.nl/pub/76075

Gharehgozli, A.H., *Developing New Methods for Efficient Container Stacking Operations*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2012-269-LIS, http://repub.eur.nl/pub/37779

Gils, S. van, *Morality in Interactions: On the Display of Moral Behavior by Leaders andEmployees*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2012-270-ORG, http://repub.eur.nl/pub/38027

Ginkel-Bieshaar, M.N.G. van, *The Impact of Abstract versus Concrete Product Communications on Consumer Decision-making Processes*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-256-MKT, http://repub.eur.nl/pub/31913

Gkougkousi, X., *Empirical Studies in Financial Accounting*, Promotor(s): Prof.dr. G.M.H. Mertens & Prof.dr. E. Peek, EPS-2012-264-F&A, http://repub.eur.nl/pub/37170

Glorie, K.M., *Clearing Barter Exchange Markets: Kidney Exchange and Beyond*, Promotor(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. J.J. van de Klundert, EPS-2014-329-LIS, http://repub.eur.nl/pub/77183

Hekimoglu, M., *Spare Parts Management of Aging Capital Products,* Promotor: Prof.dr.ir. R. Dekker, EPS-2015-368-LIS, http://hdl.handle.net/1765/79092

Heij, C.V., *Innovating beyond Technology. Studies on how management innovation, co-creation and business model innovation contribute to firm's (innovation) performance*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-370-STR, http://repub.eur.nl/pub/78651

Heyde Fernandes, D. von der, *The Functions and Dysfunctions of Reminders*, Promotor(s): Prof.dr. S.M.J. van Osselaer, EPS-2013-295-MKT, http://repub.eur.nl/pub/41514

Heyden, M.L.M., *Essays on Upper Echelons & Strategic Renewal: A Multilevel Contingency Approach*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-259-STR, http://repub.eur.nl/pub/32167

Hoever, I.J., *Diversity and Creativity*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2012-267-ORG, http://repub.eur.nl/pub/37392

Hogenboom, A.C., *Sentiment Analysis of Text Guided by Semantics and Structure,* Promotor(s):Prof.dr.ir. U.Kaymak & Prof.dr. F.M.G. de Jong, EPS-2015-369-LIS, http://hdl.handle.net/1765/79034

Hogenboom, F.P., *Automated Detection of Financial Events in News Text*, Promotor(s): Prof.dr.ir. U. Kaymak & Prof.dr. F.M.G. de Jong, EPS-2014-326-LIS, http://repub.eur.nl/pub/77237

Hollen, R.M.A.*, Exploratory Studies into Strategies to Enhance Innovation-Driven International Competitiveness in a Port Context: Toward Ambidextrous Ports,* Promotor(s) Prof.dr.ing. F.A.J. Van Den Bosch & Prof.dr. H.W.Volberda, EPS-2015-372-S&E, hdl.handle.net/1765/78881

Hoogendoorn, B., *Social Entrepreneurship in the Modern Economy: Warm Glow, Cold Feet*, Promotor(s): Prof.dr. H.P.G. Pennings & Prof.dr. A.R. Thurik, EPS-2011-246-STR, http://repub.eur.nl/pub/26447

Hoogervorst, N., *On The Psychology of Displaying Ethical Leadership: A Behavioral Ethics Approach*, Promotor(s): Prof.dr. D. de Cremer & Dr. M. van Dijke, EPS-2011-244-ORG, http://repub.eur.nl/pub/26228

Hout, D.H. van, *Measuring Meaningful Differences: Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling*, Promotor(s): Prof.dr. P.J.F. Groenen & Prof.dr. G.B. Dijksterhuis, EPS-2014-304-MKT, http://repub.eur.nl/pub/50387

Houwelingen, G.G. van, *Something To Rely On*, Promotor(s): Prof.dr. D. de Cremer & Prof.dr. M.H. van Dijke, EPS-2014-335-ORG, http://repub.eur.nl/pub/77320

Hurk, E. van der, *Passengers, Information, and Disruptions*, Promotor(s): Prof.dr. L.G. Kroon & Prof.mr.dr. P.H.M. Vervest, EPS-2015-345-LIS, http://repub.eur.nl/pub/78275

Hytonen, K.A., *Context Effects in Valuation, Judgment and Choice: A NeuroscientificApproach*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2011-252-MKT, http://repub.eur.nl/pub/30668

Iseger, P. den, *Fourier and Laplace Transform Inversion with Applications in Finance*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2014-322-LIS, http://repub.eur.nl/pub/76954

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2013-288-LIS, http://repub.eur.nl/pub/39933

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2011-222-LIS, http://repub.eur.nl/pub/22156

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promotor(s): Prof.dr. S. Stremersch, EPS-2011-239-MKT, http://repub.eur.nl/pub/23610

Karreman, B., *Financial Services and Emerging Markets*, Promotor(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, http://repub.eur.nl/pub/22280

Khanagha, S., *Dynamic Capabilities for Managing Emerging Technologies*, Promotor(s): Prof.dr. H.W. Volberda, EPS-2014-339-S&E, http://repub.eur.nl/pub/77319

Kil, J., *Acquisitions Through a Behavioral and Real Options Lens*, Promotor(s): Prof.dr. H.T.J. Smit, EPS-2013-298-F&A, http://repub.eur.nl/pub/50142

Klooster, E. van ’t, *Travel to Learn: the Influence of Cultural Distance on Competence Development in Educational Travel*, Promotor(s): Prof.dr. F.M. Go & Prof.dr. P.J. van Baalen, EPS-2014-312-MKT, http://repub.eur.nl/pub/51462

Koendjbiharie, S.R., *The Information-Based View on Business Network Performance: Revealing the Performance of Interorganizational Networks*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.mr.dr. P.H.M. Vervest, EPS-2014-315-LIS, http://repub.eur.nl/pub/51751

Koning, M., *The Financial Reporting Environment: The Role of the Media, Regulators and Auditors*, Promotor(s): Prof.dr. G.M.H. Mertens & Prof.dr. P.G.J. Roosenboom, EPS-2014-330-F&A, http://repub.eur.nl/pub/77154

Konter, D.J., *Crossing Borders with HRM: An Inquiry of the Influence of Contextual Differences in the Adoption and Effectiveness of HRM*, Promotor(s): Prof.dr. J. Paauwe & Dr. L.H. Hoeksema, EPS-2014-305-ORG, http://repub.eur.nl/pub/50388

Korkmaz, E., *Bridging Models and Business: Understanding Heterogeneity in HiddenDrivers of Customer Purchase Behavior*, Promotor(s): Prof.dr. S.L. van de Velde & Prof.dr. D. Fok, EPS-2014-316-LIS, http://repub.eur.nl/pub/76008

Kroezen, J.J., *The Renewal of Mature Industries: An Examination of the Revival of the Dutch Beer Brewing Industry*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2014-333-S&E, http://repub.eur.nl/pub/77042

Kysucky, V., *Access to Finance in a Cros-Country Context*, Promotor(s): Prof.dr. L. Norden, EPS-2015-350-F&A, http://repub.eur.nl/pub/78225

Lam, K.Y., *Reliability and Rankings*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-230-MKT, http://repub.eur.nl/pub/22977

Lander, M.W., *Profits or Professionalism? On Designing Professional Service Firms*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2012-253- ORG, http://repub.eur.nl/pub/30682

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promotor(s): Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT, http://repub.eur.nl/pub/23504

Lee, C.I.S.G, *Big Data in Management Research: Exploring New Avenues*, Promotor(s): Prof.dr. S.J. Magala & Dr W.A. Felps, EPS-2016-365-ORG, http://repub.eur.nl/pub/79818

Legault-Tremblay, P.O., Corporate Governance During Market Transition: Heterogeneous responses to Institution Tensions in China, Promotor: Prof.dr. B. Krug, EPS-2015-362-ORG, http://repub.eur.nl/pub/78649

Lenoir, A.S. *Are You Talking to Me? Addressing Consumers in a Globalised World,* Promotor(s) Prof.dr. S. Puntoni & Prof.dr. S.M.J. van Osselaer, EPS-2015-363-MKT, http://hdl.handle.net/1765/79036

Leunissen, J.M., *All Apologies: On the Willingness of Perpetrators to Apologize*, Promotor(s): Prof.dr. D. de Cremer & Dr. M. van Dijke, EPS-2014-301-ORG, http://repub.eur.nl/pub/50318

Li, D., *Supply Chain Contracting for After-sales Service and Product Support*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2015-347-LIS, http://repub.eur.nl/pub/78526

Li, Z., *Irrationality: What, Why and How*, Promotor(s): Prof.dr. H. Bleichrodt, Prof.dr. P.P. Wakker, & Prof.dr. K.I.M. Rohde, EPS-2014-338-MKT, http://repub.eur.nl/pub/77205

Liang, Q.X., *Governance, CEO Identity, and Quality Provision of Farmer Cooperatives*, Promotor(s): Prof.dr. G.W.J. Hendrikse, EPS-2013-281-ORG, http://repub.eur.nl/pub/39253

Liket, K., *Why 'Doing Good' is not Good Enough: Essays on Social Impact Measurement*, Promotor(s): Prof.dr. H.R. Commandeur & Dr. K.E.H. Maas, EPS-2014-307-STR, http://repub.eur.nl/pub/51130

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research*, Promotor(s): Prof.dr. A.R. Thurik, Prof.dr. P.J.F. Groenen, & Prof.dr. A. Hofman, EPS-2013-287-S&E, http://repub.eur.nl/pub/40081

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promotor(s): Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, http://repub.eur.nl/pub/22814

Lu, Y., *Data-Driven Decision Making in Auction Markets*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. W. Ketter, EPS-2014-314-LIS, http://repub.eur.nl/pub/51543

Manders, B., *Implementation and Impact of ISO 9001*, Promotor(s): Prof.dr. K. Blind, EPS-2014-337-LIS, http://repub.eur.nl/pub/77412

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promotor(s): Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, http://repub.eur.nl/pub/22744

Mees, H., *Changing Fortunes: How China's Boom Caused the Financial Crisis*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2012-266-MKT, http://repub.eur.nl/pub/34930

Mell, J.N., *Connecting Minds: On The Role of Metaknowledge in Knowledge Coordination*, Promotor: Prof.dr.D.L. van Knippenberg, EPS-2015-359-ORG, http://hdl.handle.net/1765/78951

Meuer, J., *Configurations of Inter-firm Relations in Management Innovation: A Study in China's Biopharmaceutical Industry*, Promotor(s): Prof.dr. B. Krug, EPS-2011-228-ORG, http://repub.eur.nl/pub/22745

Micheli, M.R., *Business Model Innovation: A Journey across Managers' Attention and Inter-Organizational Networks*, Promotor(s): Prof.dr. J.J.P. Jansen, EPS-2015-344-S&E, http://repub.eur.nl/pub/78241

Mihalache, O.R., *Stimulating Firm Innovativeness: Probing the Interrelations between Managerial and Organizational Determinants*, Promotor(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch, & Prof.dr. H.W. Volberda, EPS-2012-260-S&E, http://repub.eur.nl/pub/32343

Milea, V., *News Analytics for Financial Decision Support*, Promotor(s): Prof.dr.ir. U. Kaymak, EPS-2013-275-LIS, http://repub.eur.nl/pub/38673

Naumovska, I., *Socially Situated Financial Markets: A Neo-Behavioral Perspective on Firms, Investors and Practices*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. A. de Jong, EPS-2014-319-S&E, http://repub.eur.nl/pub/76084

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in short term planning and in disruption management*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2011-224-LIS, http://repub.eur.nl/pub/22444

Nuijten, A.L.P., *Deaf Effect for Risk Warnings: A Causal Examination applied to Information Systems Projects*, Promotor(s): Prof.dr. G.J. van der Pijl, Prof.dr. H.R. Commandeur & Prof.dr. M. Keil, EPS-2012-263-S&E, http://repub.eur.nl/pub/34928

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on bureaucracy and formal rules*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, http://repub.eur.nl/pub/23250

Ozdemir, M.N., *Project-level Governance, Monetary Incentives, and Performance in Strategic R&D Alliances*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, http://repub.eur.nl/pub/23550

Peers, Y., *Econometric Advances in Diffusion Models*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-251-MKT, http://repub.eur.nl/pub/30586

Peters, M., *Machine Learning Algorithms for Smart Electricity Markets*, Promotor(s): Prof.dr. W. Ketter, EPS-2014-332-LIS, http://repub.eur.nl/pub/77413

Porck, J., *No Team is an Island: An Integrative View of Strategic Consensus between Groups*, Promotor(s): Prof.dr. P.J.F. Groenen & Prof.dr. D.L. van Knippenberg, EPS-2013-299-ORG, http://repub.eur.nl/pub/50141

Porras Prado, M., *The Long and Short Side of Real Estate, Real Estate Stocks, and Equity*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2012-254-F&A, http://repub.eur.nl/pub/30848

Poruthiyil, P.V., *Steering Through: How organizations negotiate permanent uncertaintyand unresolvable choices*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. S.J. Magala, EPS-2011-245-ORG, http://repub.eur.nl/pub/26392

Pourakbar, M., *End-of-Life Inventory Decisions of Service Parts*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2011-249-LIS, http://repub.eur.nl/pub/30584

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promotor(s): Prof.dr. H.J.H.M. Claassen & Prof.dr. H.R. Commandeur, EPS-2013-282-S&E, http://repub.eur.nl/pub/39654

Protzner, S. *Mind the gap between demand and supply: A behavioral perspective on demand forecasting,* Promotor(s): Prof.dr. S.L. van de Velde & Dr. L. Rook, EPS-2015-364-LIS, *http*://repub.eur.nl/pub/79355

Pruijssers, J.K., *An Organizational Perspective on Auditor Conduct*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2015-342-S&E, http://repub.eur.nl/pub/78192

Retel Helmrich, M.J., *Green Lot-Sizing*, Promotor(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-291-LIS, http://repub.eur.nl/pub/41330

Rietdijk, W.J.R. *The Use of Cognitive Factors for Explaining Entrepreneurship,* Promotor(s): Prof.dr. A.R. Thurik & Prof.dr. I.H.A. Franken, EPS-2015-356-S&E, http://repub.eur.nl/pub/79817

Rietveld, N., *Essays on the Intersection of Economics and Biology*, Promotor(s): Prof.dr. A.R. Thurik, Prof.dr. Ph.D. Koellinger, Prof.dr. P.J.F. Groenen, & Prof.dr. A. Hofman, EPS-2014-320-S&E, http://repub.eur.nl/pub/76907

Rijsenbilt, J.A., *CEO Narcissism: Measurement and Impact*, Promotor(s): Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, http://repub.eur.nl/pub/23554

Rösch, D. *Market Efficiency and Liquidity,* Promotor: Prof.dr. M.A. van Dijk, EPS-2015-353-F&A, http://hdl.handle.net/1765/79121

Roza-van Vuren, M.W., *The Relationship between Offshoring Strategies and Firm Performance: Impact of innovation, absorptive capacity and firm size*, Promotor(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, http://repub.eur.nl/pub/22155

Rubbaniy, G., *Investment Behaviour of Institutional Investors*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2013-284-F&A, http://repub.eur.nl/pub/40068

Schoonees, P. *Methods for Modelling Response Styles*, Promotor: Prof.dr P.J.F. Groenen, EPS-2015-348-MKT, http://repub.eur.nl/pub/79327

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promotor(s): Prof.dr. G.M.H. Mertens, EPS-2013-283-F&A, http://repub.eur.nl/pub/39655

Sousa, M.J.C. de, *Servant Leadership to the Test: New Perspectives and Insights*, Promotor(s): Prof.dr. D.L. van Knippenberg & Dr. D. van Dierendonck, EPS-2014-313-ORG, http://repub.eur.nl/pub/51537

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2013-293-LIS, http://repub.eur.nl/pub/41513

Staadt, J.L., *Leading Public Housing Organisation in a Problematic Situation: A Critical Soft Systems Methodology Approach*, Promotor(s): Prof.dr. S.J. Magala, EPS-2014-308-ORG, http://repub.eur.nl/pub/50712

Stallen, M., *Social Context Effects on Decision-Making: A Neurobiological Approach*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2013-285-MKT, http://repub.eur.nl/pub/39931

Tarakci, M., *Behavioral Strategy: Strategic Consensus, Power and Networks*, Promotor(s): Prof.dr. D.L. van Knippenberg & Prof.dr. P.J.F. Groenen, EPS-2013-280-ORG, http://repub.eur.nl/pub/39130

Teixeira de Vasconcelos, M., *Agency Costs, Firm Value, and Corporate Investment*, Promotor(s): Prof.dr. P.G.J. Roosenboom, EPS-2012-265-F&A, http://repub.eur.nl/pub/37265

Troster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://repub.eur.nl/pub/23298

Tsekouras, D., *No Pain No Gain: The Beneficial Role of Consumer Effort in Decision-Making*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-268-MKT, http://repub.eur.nl/pub/37542

Tuijl, E. van, *Upgrading across Organisational and Geographical Configurations*, Promotor(s): Prof.dr. L. van den Berg, EPS-2015-349-S&E, http://repub.eur.nl/pub/78224

Tuncdogan, A., *Decision Making and Behavioral Strategy: The Role of Regulatory Focus in Corporate Innovation Processes*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch, Prof.dr. H.W. Volberda, & Prof.dr. T.J.M. Mom, EPS-2014-334-S&E, http://repub.eur.nl/pub/76978

Uijl, S. den, *The Emergence of De-facto Standards*, Promotor(s): Prof.dr. K. Blind, EPS-2014-328-LIS, http://repub.eur.nl/pub/77382

Vagias, D., *Liquidity, Investors and International Capital Markets*, Promotor(s): Prof.dr. M.A. van Dijk, EPS-2013-294-F&A, http://repub.eur.nl/pub/41511

Veelenturf, L.P., *Disruption Management in Passenger Railways: Models for Timetable, Rolling Stock and Crew Rescheduling*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2014-327-LIS, http://repub.eur.nl/pub/77155

Venus, M., *Demystifying Visionary Leadership: In search of the essence of effective vision communication*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-289-ORG, http://repub.eur.nl/pub/40079

Vermeer, W., *Propagation in Networks:The impact of information processing at the actor level on system-wide propagation dynamics,* Promotor: Prof.mr.dr. P.H.M.Vervest, EPS-2015-373-LIS, http://repub.eur.nl/pub/79325

Visser, V.A., *Leader Affect and Leadership Effectiveness:How leader affective displays influence follower outcomes*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-286-ORG, http://repub.eur.nl/pub/40076

Vlam, A.J., *Customer First? The Relationship between Advisors and Consumers of Financial Products*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-250-MKT, http://repub.eur.nl/pub/30585

Vries, J. de, *Behavioral Operations in Logistics,* Promotor(s): Prof.dr M.B.M de Koster & Prof.dr. D.A. Stam, EPS-2015-374-LIS, http://repub.eur.nl/pub/79705

Waltman, L., *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*, Promotor(s): Prof.dr.ir. R. Dekker & Prof.dr.ir. U. Kaymak, EPS-2011-248- LIS, http://repub.eur.nl/pub/26564

Wang, T., *Essays in Banking and Corporate Finance*, Promotor(s): Prof.dr. L. Norden & Prof.dr. P.G.J. Roosenboom, EPS-2015-352-F&A, http://repub.eur.nl/pub/78301

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, http://repub.eur.nl/pub/26066

Wang, Y., *Corporate Reputation Management: Reaching Out to Financial Stakeholders*, Promotor(s): Prof.dr. C.B.M. van Riel, EPS-2013-271-ORG, http://repub.eur.nl/pub/38675

Weenen, T.C., *On the Origin and Development of the Medical Nutrition Industry*, Promotor(s): Prof.dr. H.R. Commandeur & Prof.dr. H.J.H.M. Claassen, EPS-2014-309-S&E, http://repub.eur.nl/pub/51134

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value*, Promotor(s): Prof.dr. A. de Jong, EPS-2013-277-F&A, http://repub.eur.nl/pub/39127

Yang, S., *Information Aggregation Efficiency of Prediction Markets*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck, EPS-2014-323-LIS, http://repub.eur.nl/pub/77184

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2013-276-LIS, http://repub.eur.nl/pub/38766

Zhang, D., *Essays in Executive Compensation*, Promotor(s): Prof.dr. I. Dittmann, EPS- 2012-261-F&A, http://repub.eur.nl/pub/32344

Zwan, P.W. van der, *The Entrepreneurial Process: An International Analysis of Entry and Exit*, Promotor(s): Prof.dr. A.R. Thurik & Prof.dr. P.J.F. Groenen, EPS-2011-234-ORG, http://repub.eur.nl/pub/23422

Erasmus Research Institute of Management - ERiM

TEXTUAL ANALYSIS OF INTANGIBLE INFORMATION

Traditionally, equity investors have relied upon the information reported in firms' financial accounts to make their investment decisions. Due to the conservative nature of accounting standards, firms cannot value their intangible assets such as corporate culture, brand value and reputation. Investors' efforts to collect such information have been hampered by the voluntary nature of Corporate Social Responsibility (CSR) reporting standards, which have resulted in the publication of inconsistent, stale and incomplete information across firms. To address this deficiency, this thesis investigates the problem of designing automated approaches to infer measures of intangible information from a firm's stakeholders (namely its employees, investors, the media, and regulators) using Web data. In contrast to accounting data which reside in a traditional row-column database, Web data are considered "unstructured". This is because the variety of text and multimedia content available on the Web doesn't fit neatly into a structured database. The first three studies in this thesis are methodological and draw upon techniques commonly adopted in the fields of Natural Language Processing (NLP) and Information Retrieval to infer intangible information from text. The second three studies draw upon techniques from financial asset pricing literature and investigate how intangible information may be integrated into financial statement analysis. Our findings highlight the merits of exploiting unstructured data for investors seeking to integrate a wider variety of information into their investment decisions and processes.

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERIM PhD Series
Research in Management

Erasmus Research Institute of Management - ERiM
Rotterdam School of Management (RSM)
Erasmus School of Economics (ESE)
Erasmus University Rotterdam (EUR)
P.O. Box 1738, 3000 DR Rotterdam,
The Netherlands

Tel. +31 10 408 11 82
Fax +31 10 408 96 40
E-mail info@erim.eur.nl
Internet www.erim.eur.nl

Design & layout: B&T Ontwerp en advies (www.b-en-t.nl) Print: Haveka (www.haveka.nl)