Master's thesis

2020

Peder Gjerstad, Peter Filip Meyn, Thomas Dowling Næss

Master's thesis

NTNU
Norwegian University of
Science and Technology
Faculty of Economics and Management
Department of Industrial Economics and Technology
Management

Peder Gjerstad
Peter Filip Meyn
Thomas Dowling Næss

Stock Market Predictions Using Advanced Textual Analysis of Annual Reports

June 2020

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Stock Market Predictions Using Advanced Textual Analysis of Annual Reports

Peder Gjerstad
Peter Filip Meyn
Thomas Dowling Næss

Industrial Economics and Technology Management
Submission date: June 2020
Supervisor: Peter Molnár

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

Abstract

Are investors and analysts effective in interpreting the content in annual reports? This thesis suggests that financial markets in the short term overlook important information contained in the annual reports and that it is possible to use techniques from Natural Language Processing (NLP) on the annual reports to generate useful input when valuing listed companies. We analyze $15,700$ annual reports published by S&P 500 companies in the period from 1994 to 2018 and find that one-year abnormal return decreases significantly with the amount of negative sentiment in reports and with the reports' file size. Interestingly, the effects are not reflected in stock prices several days after reports are published, suggesting that is takes a long time for the market to absorb this type of information. Through the use of Latent Dirichlet Allocation (LDA), we find that annual reports with a focus on ”Health care”, ”Environmental cost” and ”Financial plans” in their forward-looking statements tend to see higher abnormal returns, while focus on ”Lawsuits”, ”Property lease” and ”Foreign exchange” precede negative abnormal returns. Finally, a trading strategy based on sentiment, readability, and topics addressed in annual reports generate an annualized risk-adjusted return of 3.8% on an out-of-sample dataset from 2004 to 2018.

Sammendrag

Får investorer og analytikere med seg all informasjon som finnes i selskapers årsrapporter? I denne oppgaven finner vi at finansmarkedene på kort sikt overser viktig informasjon som er inneholdt i årsrapportene, og at denne informasjonen kan utnyttes gjennom teknikker for behandling av naturlig språk for å utgjøre en del av grunnlaget i verdsettelsen av børsnoterte selskaper. Vi analyserer $15,700$ årsrapporter publisert av selskaper på aksjeindeksen S&P 500 i perioden 1994 til 2018, og finner at ettårig abnormal avkastning synker signifikant med mengden negativt sentiment i rapportene og med rapportenes filstørrelse. Vi finner derimot ikke slike mønstre på kort sikt, noe som kan indikere at det tar lang tid før finansmarkedene plukker opp denne typen informasjon. Ved å bruke Latent Dirichlet Allocation (LDA), finner vi at årsrapporter som fokuserer på ”Helsetjenester”, ”Miljøkostnader” og ”Finansielle planer” assosieres med høyere abnormal avkastning, mens økt fokus på ”Søksmål”, ”Eiendomskontrakter” og ”Utenlandsk valuta” ofte assosieres med det motsatte. Til slutt lager vi en modell for aksjehandel basert på metaalgoritmen AdaBoost og CART beslutningstrær for å vise at kun ved å benytte informasjon om sentiment, lesbarhet og temaer som rapporten tar opp, er det mulig å generere en årlig risikojustert avkastning på 3.8% på et datasett fra 2004 til 2018.

Preface

This thesis is submitted in partial fulfilment of the requirements for the awards of Master of Science in Industrial Economics and Technology Management, with specializations in finance.

We wish to extend our most sincere gratitude to our supervisor, Associate Professor Peter Molnár, for his genuine interest in our work, unprecedented flexibility, and sound advice.

Contents

List of Tables

List of Figures

Abbreviations

AAER Accounting and Auditing Enforcement Release. 8

ANEW The Affective Norms for English Words. 6, 63

CIK Central Index Keys. 12

CRSP Center for Research in Security Prices. 11

CUSIP Committee on Uniform Securities Identification Procedures. 12

EDGAR Electronic Data Gathering, Analysis and Retrieval. 4, 11, 17

EW Equally Weighted. 43, 45

FLS Forward-Looking Statements. 13, 17

GI General Inquirer. 4, 5

LDA Latent Dirichlet Allocation. 3, 4, 8, 9, 32, 35, 36, 38, 40, 52, 71

LIWC Linguistic Inquiry and Word Count. 6, 63

MD&A Management's Discussion and Analysis of Financial Position and Results of Operations. 5, 8, 9, 13

NLP Natural Language Processing. 1, 3, 4, 8, 14, 31

NPMI Normalized Pointwise Mutual Information. 35

PMI Pointwise Mutual Information. 35

SEC U.S Securities and Exchange Commission. 4, 6, 8, 11, 31, 32

TF-IDF Term Frequency - Inverse Document Frequency. 14, 15, 19, 26

VADER Valance Aware Dictionary for Sentiment Reasoning. 2, 6, 14, 19, 26

WRDS Wharton Research Data Services. 11, 12

Chapter 1

Introduction

In this paper, we use methods from Natural Language Processing (NLP) to analyze the qualitative content of $15,700$ annual reports published by S&P-500 companies in the period between 1994 and 2018. We focus on three main aspects of the reports; the language sentiment, the readability, and the topics that the reports are discussing, and we analyze how this relates to subsequent financial return, trading volume and volatility after the reports are published. We further use the findings to build a trading strategy that achieves an abnormal return of 3.8% annually in an out-of-sample dataset.

Investors have numerous available information sources for valuing publicly traded stocks, such as financial statements, news articles, and earnings calls. The information sources contain both quantitative and qualitative information, but investors and analysts have historically focused almost exclusively on the former. Qualitative information is generally less precise than quantitative information (McDonald and Loughran, 2015). McDonald and Loughran (2015) also argues that quantitative research is more advanced and has well-established norms making it easier to compare work done (e.g. valuations) by different practitioners.

However, the increasing amount of available online information and the expanding growth in computational power has raised the attention on textual analysis in recent years. Although textual analysis in its primary forms dates back to the 1300s, almost all studies on the topic of accounting and finance have been published the past few decades (McDonald and Loughran, 2015). At the same time, the rise in the amount of information produced makes the search and processing of textual data complex. If investors are unable to keep up with the increasing magnitude and complexity of this data, disclosed information may go unattended (Cohen et al., 2018).

Several studies have shown significant relations between information extracted with textual analysis methods and market reactions (Li, 2006; Tetlock, 2007; Jiang et al., 2019). Karapandza (2016) is the first to establish a link between textual data in annual reports and long-run changes in stock prices, by showing that companies that talk less about the future generate significant positive abnormal returns of about 5% per year. Also, Loughran and McDonald (2011) find significant short term stock price effects based on the level of negative sentiment in the annual reports but do not find statistically significant results for the long run.

Among the vast amount of information sources, we focus our textual analysis on 10-K filings, as these annual reports are audited, unlike other types of financial reports from corporations. We include annual reports from all companies that have been listed on the S&P-500 stock index during the period 1994 - 2018. We then extract qualitative information, such as the reports' sentiment, the reports' readability, and the topics the reports discuss. Generally, 10-K reports do not contain accounting information not already known to investors. This information can be deduced from the quarterly reports, the last of which should be published no more than 30 days (about 22 trading days) in advance of the 10-K annual report. Hence, controlling for accounting information like standardized unexpected earnings is not warranted.

We find that more negative language sentiment in the report is associated with lower abnormal returns the following year. Furthermore, by using the natural logarithm of the filesize as a proxy for readability, we find that abnormal return decreases with the size of the report, while volatility and trading volume increases with report size. We do not, however, find statistically significant patterns between short-term abnormal returns and either sentiment or readability, indicating that financial markets are ineffective in absorbing the textual information contained in the reports.

In the topic analysis, we find that abnormal return increases with the amount of discussions about ”Health care”, ”Environmental cost”, and ”Financial plans” and decreases with increased discussions about ”Lawsuits”, ”Property lease”, and ”Foreign exchange”. These results are, however, less significant than the results for sentiment and readability.

We use the findings in a trading strategy context, where the meta-algorithm AdaBoost with 200 CART decision trees leverages the annual reports' topics and sentiment to generate an annualized risk-adjusted return of 3.8% on an out-of-sample dataset from 2004 to 2018.

This thesis contributes to the literature in two important ways. To the best of our knowledge, we are the first to systematically identify relationships between the topics addressed in a corporate annual report, and subsequent financial performance. Secondly, we show that our novel measure combining the best available word list tailored to company annual reports with the sophisticated sentiment analysis tool, Valance Aware Dictionary for Sentiment Reasoning (VADER), is superior to the conventional measures based on the bag-of-words assumption.

The thesis is structured as follows. In Chapter 2, we discuss existing research related to our thesis, and in Chapter 3 we present the data sources, explain the data cleaning and data preprocessing performed, and define the variables we use. Chapter 4 presents the results for the analysis on readability and sentiment, while Chapter 5 contains the results from the analysis based on topics. Chapter 6 explains the methodology and results for the simulated trading, before we conclude in Chapter 7.

Chapter 2

Overview of Textual Analysis in Finance

To computationally analyze the effects of 10-K filings, we have to make use of textual analysis techniques commonly referred to as Natural Language Processing (NLP). There are two main approaches to analyze natural language; lexicon-based and machine learning (Guo et al., 2016). Figure 2.1 illustrates a simplified overview of the most common textual classification methods used within accounting and finance, in addition to Latent Dirichlet Allocation (LDA). Whereas lexicon-based techniques are dependent on dictionaries where humans have assigned values to a set of words, machine learning techniques could either be supervised, in which case it is based on human-provided target values (typically for the text as a whole), or it could be unsupervised learning, requiring no human input. In this paper, we use combinations of readability, dictionary-based sentiment analysis, and topic modeling with LDA. Existing literature covering these three methods will be discussed in this chapter. For insights on the remaining methods, we refer to Kearney and Liu (2014); Kumar and Ravi (2016); Guo et al. (2016); Loughran and McDonald (2016, 2019).

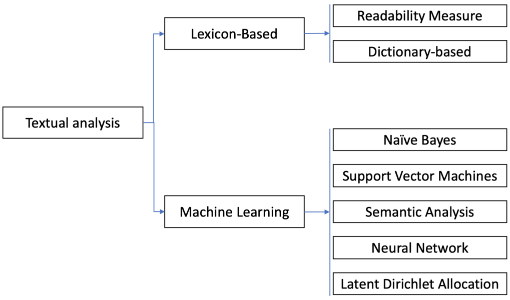


Figure 2.1: Textual Analysis

2.1 10-K Filings

The 10-K filings, or annual reports, mandated by U.S Securities and Exchange Commission (SEC) have historically contained little vital information that is not already known by the investors either through previous earnings releases or company conference calls. Therefore, it would be reasonable to assume that the filing itself has little or no significant market impact, as were the results of early empirical research (Easton and Zmijewski, 1993). However, more recent studies (Asthana and Balsam, 2001; You and Zhang, 2009; Karapandza, 2016) indicate that the 10-K filings indeed do impact all of the companies return, volatility, and volume. There are several plausible explanations for the discrepancies in these findings. Dyer et al. (2017) show, with the use of LDA, that the disclosures are getting both longer and more complex, containing more information that might be useful for the investors, but also becoming more challenging to comprehend. Perhaps more importantly, the accessibility of disclosures, especially for smaller investors, has increased considerably after the implementation of SEC's Electronic Data Gathering, Analysis and Retrieval (EDGAR). Whereas Easton and Zmijewski (1993) showed that pre-EDGAR filings caused no significant market reactions, Asthana and Balsam (2001) find that short-term market reactions to 10-K filings after implementation of the new filing system are significant both in terms of higher trading volume and positive abnormal returns, and that they differ from the reactions caused by the earlier filings.

2.2 Sentiment Analysis

We define sentiment analysis for our purpose as the process of using NLP to systematically extract information about the polarity of any expressed opinion in a text.

2.2.1 Traditional Lexicons used within Accounting and Finance

In this section we will highlight and compare the most relevant sentiment analysis methods and measures used within the dictionary-based approach. As discussed in Loughran and McDonald (2016) there are four dictionaries that have dominated research within accounting and finance:

- Harvard General Inquirer (GI)[1]
- Diction[2]
- Henry (2008)
- Loughran and McDonald (2011)

Harvard GI (specifically the Harvard-IV-4 TagNeg) and Diction were not intentionally made for accounting and financial purposes, but have been frequently used because they have been easily accessible. Tetlock (2007), one of the most prominent papers on sentiment analysis within accounting and finance, uses Harvard GI to examine the relationship

[1] Latest version available through: http://www.wjh.harvard.edu/ inquirer/homecat.htm

[2] 35 different dictionary subcategories available through: http://www.dictionsoftware.com

between The Wall Street Journal's column ”Abreast of the Market” and stock market returns. He finds that high media pessimism results in low subsequent stock returns and that unusually high or low pessimism results in higher trading volumes. Despite several other researchers successfully employing either of the Harvard GI and Diction dictionaries in capturing sentiment tone (Tetlock et al., 2008; Davis et al., 2011; Rogers et al., 2011; Davis and Tama-Sweet, 2011), Loughran and McDonald (2011) criticize both Harvard GI and Diction for failing to capture the managerial tone in 10-K filings. They justify this by showing that 75% of the negative words in Harvard's GI do not necessarily have a pessimistic meaning in corporate filings, such as *tax*, *depreciation* and *capital*. Loughran and McDonald (2015) find similar results for Diction. Loughran and McDonald (2016) further support this criticism by referring to the work of Li (2010), who uses both dictionaries and finds no relation between future stock performance and the tone in the Management's Discussion and Analysis of Financial Position and Results of Operations (MD&A) section of 10-K filings.

The Henry (2008) dictionary was created using earnings press releases, and is, according to Loughran and McDonald (2016), most likely the first dictionary made intentionally for financial documents. Price et al. (2012) used the dictionary to show that stock returns were significantly higher after conference calls with a positive tone in the Q&A session, and significantly lower when the tone was negative. They also report that Harvard GI provides less significant results, and thus that using Henry (2008) is more appropriate for analyzing business conference calls. Doran et al. (2012) also find that stock returns are significantly correlated with the tone in conference calls. However, the Henry (2008) is comprised of a small sample of words (only 85 negative and 105 positive), consequently limiting the applicability and effectiveness of the dictionary.

To counter the challenges related to the traditional word lists, Loughran and McDonald (2011) constructed several word lists containing words that are classified as positive, negative, uncertain, litigious, strong modal, and weak modal in the context of 10-K filings. Results from their work show that the negative word list performs better than Harvard GI in capturing the tone of 10-K filings and that this dictionary can be used to predict announcement returns. Davis et al. (2014) use several wordlists namely Henry (2008), Diction, and Loughran and McDonald (2011) to evaluate the impact of manager-specific optimism of the tone used during earning conference calls. They find significant results when using Henry (2008) and Loughran and McDonald (2011), but not with Diction, illustrating that the ”financial” dictionaries seem to be more appropriate in analyzing business disclosures. Jiang et al. (2019) use the Loughran and McDonald (2011) word lists to create a manager sentiment index based on 10-K and 10-Q filings. They show that a one-standard-deviation increase in manager sentiment is associated with a 1.26% decrease in the expected excess market return for the next month, i.e., that high manager sentiment tends to predict lower future stock returns. Also, their results indicate that the index has greater predictive power than other macroeconomic variables, demonstrating that the dictionary is efficient in capturing sentiment tone of 10-K filings.

While Loughran and McDonald (2011) highlight the importance of dictionary choice, the authors do not find a statistically significant relationship between the tone of a 10-K report and subsequent long-run abnormal returns.

2.2.2 Sentiment Analysis using VADER

VADER is a parsimonious rule-based model for sentiment analysis developed by Hutto and Gilbert (2014). The VADER sentiment lexicon was created by first gathering 9,000 lexical feature *candidates* into a list. The feature candidates were gathered from three already approved lexicons; the Harvard GI, Linguistic Inquiry and Word Count (LIWC), and The Affective Norms for English Words (ANEW)[3]. In addition, the list includes several other sentiment expressions often used in social media, such as abbreviations, slang, and emoticons. A sentiment valance score (intensity), ranging between -4 (most negative) and 4 (most positive), is then generated for each feature in the list by using the wisdom-of-crowd method with ten human raters. Features with a non-zero mean and a standard deviation of less than 2.5 were kept in the list, while the others were excluded. The final VADER sentiment lexicon consists of approximately 7,500 features. In addition to this lexicon, VADER also consists of a set of heuristic rules which were defined by analyzing text from 800 tweets. The goal of the evaluation was to find textual attributes that affected the perceived sentiment intensity. Most importantly, the rules enable capturing changes in sentiment intensity based on the syntactic arrangement. The rules modify the score of each word in the lexicon based on degree modifiers (e.g., ”super”, ”slightly”), negation (words following e.g., ”not”), punctuation (e.g., exclamation points), capitalization (e.g., words in ALL-CAPS), and words following the contrastive conjunction ”but”.

Hutto and Gilbert (2014) stated that the intention with VADER was to create a text analyzing tool which could cope well with social media style text, but also easily generalizes to other domains. In their paper, they compared VADER to 11 well-established sentiment tools across four different domain contexts; tweets, movie reviews, technical product reviews, and N.Y. Times editorials, they found that VADER performed as well as, or better than, all of the other sentiment tools within each domain. To our knowledge, no research has so far used VADER to capture the sentiment in SEC filings, but studies have proven VADER superior in capturing sentiment on other textual domains within finance, such as financial microblogs (Sohangir et al., 2018).

VADER should, in other words, be a more sophisticated tool for sentiment analysis than straightforward counting occurrences of words from a dictionary. In this thesis, we build on this property by modifying the wisdom-of-crowd dictionary used by out-of-the-box VADER to include the positive and negative word lists collected by and contained in Loughran and McDonald (2011). By doing this, we should end up with a tool that combines the sophistication of VADER with the highly specialized domain knowledge from the 10-K-specific dictionary of Loughran and McDonald (2011). For comparison, we also implement and test a more conventional measure based solely on the Loughran and McDonald (2011) dictionary, both proportionally weighted and TF-IDF-weighted.

2.3 Readability

Readability is defined by how easily the receiver of information comprehends the intended message. The content (e.g., complexity, vocabulary, syntax) and the presentation (e.g.,

[3] We refer to the Appendix for more information on LIWC, and ANEW.

font, font size, line spacing) of the document are essential to determine the degree of readability. The traditional *Gunning Fog* (or Fog index) is one of the most popular readability measures used in linguistics. The Fog index is based on a mathematical formula created by Robert Gunning in 1952[4] (see Equation 3.5), and the index depends on the average length of sentences, the number of words and the portion of complex words (defined as words with two or more syllabuses). The existing literature on readability in business documents is extensive, and prior research dates back to the 80s. However, in early studies, sample sizes were small (Lewis et al., 1986; Tennyson et al., 1990) and results indecisive.

Li (2008) was, to our knowledge, one of the first to examine the relation between 10-K's readability and firm performance for a large sample size. He uses the Fog index and the length of documents (defined as the natural log of the number of words) as readability measures and finds that companies with annual reports with a high Fog index value tend to have lower subsequent earnings and that firms with more persistent positive earnings tend to have annual reports that are easier to read. Li (2008) suggests that the length of the documents could be used to measure disclosure complexity because firms may use longer reports to conceal damaging information strategically. He presumes that longer documents require higher information-processing costs and therefore are more challenging to read. The results provided by Li (2008) may, however, as discussed in Bloomfield (2008), be caused by poorly performing firms with the need of more sentences to explain the company situation thoroughly.

In the light of the findings of Li (2008), many researchers have continued to use the Fog index as a measure of readability. The link between investor behavior and language complexity in 10-Ks is, for instance, investigated by Miller (2010) using The Fog index and word counts as readability measures. Their findings show that firms with hard-to-read reports are associated with small investors trading relatively fewer shares close to the filing date. High Fog index value indicates that the financial documents are harder to process, and the findings are thus consistent with the results from Li (2008). Furthermore, various research combine the Fog index with other variables to link readability with actual firm performances (Hilary et al., 2009; Guay et al., 2016; Lawrence, 2013; Franco et al., 2015). Franco et al. (2015) uses Fog index, amongst others, to analyze approximately $350,000$ analyst reports to investigate investor behaviors. Their results show that easy-to-read reports are positively related to higher trading volumes around the reporting date.

The Fog index is presumably one of the most frequently implemented measures of readability, but this index has shown to be inadequate in the context of business writing. Loughran and McDonald (2014) uncover major limitations with this index and argue that the Fog index is a poor measure due to two main reasons. First and foremost, the portion of complex words (i.e., words with more than two syllabuses) in financial documents is a weak metric because investors commonly comprehend these words. For instance, Loughran and McDonald (2014) report that ”complex” words such as *company, operations, financial, period* and *management* are not likely to be confusing to a reader of financial information. Additionally, the Fog index also depends on the average number of words per sentence, which is difficult and complex to calculate accurately. Empirically, they demonstrate that the natural log of file size is a better proxy for readability, and they

[4] Plain Language at Work Newsletter [website], http://www.impact-information.com/impactinfo/newsletter/plwork08.htm, (accessed April 15, 2020)

recommend to use this approach rather than the Fog index. Even though 10-K file size may provide a solid proxy, Loughran and McDonald (2014) report that the findings cannot separate firm complexity from its written language complexity; hence, researchers should control for firm size when they consider 10-K file size in textual analysis.

Somewhat tangentially related to 10-K readability, Karapandza (2016) is the first paper to show that qualitative information in 10-K reports have systematic effects on the long-term stock performance. Karapandza (2016) uses the frequency of future tense words (e.g., *will*, *shall* and *going to*) to show that firms talking less about the future in their annual reports generate positive abnormal returns of about 5% annually.

2.4 Latent Dirichlet Allocation

Within the field of NLP, LDA, first presented in a machine learning context by Blei et al. (2003) has become a popular method to extract topics from a set of documents. Under the correct assumptions, LDA extraction should capture natural topic structures in text documents that match human interpretation (Griffiths and Steyvers, 2004; Chang et al., 2009). The approach is generally applied to text documents such as journals and articles, and the existing literature on LDA within accounting and finance is limited.

The U.S. SEC requires companies to include MD&A in their 10-K disclosures. Publicly traded firms are therefore obligated to add a narrative explanation regarding their financial statements, conditions, and operations in 10-Ks. One of the objectives is to inform the reader and improve the reader's comprehension of the current situation seen from a manager's point of view. Consequently, all firms need to consider and discuss the identical, required set of themes, and MD&As are, therefore, relatively similar across firms. As a result, some researchers use solely the MD&A sections to extract topics from 10-Ks because they are likely to yield meaningful topics (Hoberg and Lewis, 2017). The LDA model identifies topics and infers the topic distribution from an input of text documents. Even though the topics are generated automatically, the text documents chosen as input may be affected by selection biases. Hoberg and Lewis (2017) claim that using periodic disclosure platforms, such as MD&A, removes any concerns regarding proper randomization in the selection process, and they successfully identify interpretable topics from the sections. Ball et al. (2013) also extract topics discussed in MD&A sections using LDA. They use the topics to illustrate the nature of rapid change in business environments. With a benchmark of 75 generated topics and a sample of companies undergoing business change, Ball et al. (2013) identify topics that, for example, involve marketing, investment strategies, new agreements, and financial constraints. Ball et al. (2013) and Hoberg and Lewis (2017) show that LDA topic extraction from MD&As yields meaningful topics, and demonstrate that extracting the MD&A section from 10-Ks is a useful starting point for LDA modelling.

Hoberg and Lewis (2017) analyze fraudulent companies' MD&A disclosures and compare the verbal content with matched non-fraudulent firms (industry peers). The SEC issues an Accounting and Auditing Enforcement Release (AAER) when they decide to investigate a firm suspected for intentionally misrepresenting material facts. Hoberg and Lewis (2017) show that AAER firms use a vocabulary distinctive relative to peers. During the misreporting years, AAER companies are likely to publish irregular content and verbally misrepresent revenues and expenses, and, by using LDA topic extraction, they find that some topics are either under-disclosed or intentionally avoided to discuss.

Brown et al. (2017) also employ LDA on annual reports to predict misreporting, but their approach is somewhat different from the method in Hoberg and Lewis (2017). Firstly, they consider the complete report, i.e., not solely restricted to the MD&A section. They argue that although Hoberg and Lewis (2017) demonstrate that MD&A sections are appropriate for topic modelling, they miss the opportunity of capturing additional important topics written in other sections of the 10-K filing. Loughran and McDonald (2016) show that firms may exploit various sections in 10-Ks to disseminate information strategically. Hence, aiming the attention to one singular section can be a disadvantage. Secondly, Brown et al. (2017) use a sample of 10-K filings from the period 1994 - 2012 to fit their LDA model, and they apply a rolling-window analysis to capture the temporal change of language and thematic content. Brown et al. (2017), therefore, extend Hoberg and Lewis (2017) because the latter exclusively employ samples from one single year to run their LDA model and do not consider the nature of temporal variation in managerial statements. Nevertheless, the findings provided by Brown et al. (2017) show that important topics such as cost commitments and loan operations may help financiers and scientists to identify misreporting companies.

Relatively few studies investigate the relationship between topics in financial text documents and its effect on financial markets. Feuerriegel et al. (2016) employ LDA on ad hoc announcements and investigate the relationship between the generated topics and abnormal return in the German stock market. Feuerriegel et al. (2016) successfully identify significant topics that yield effects on stock prices, and their study further motivated them to do an analogous analysis on the U.S. stock market. Feuerriegel and Pröllochs (2018) use several thousands of 8-K filings to show that 5 out of 20 topics (e.g. ”Drug testing”) yield significant abnormal returns. To our knowledge, no research has considered the relationship between LDA topic extraction from 10-Ks and stock performance. In this respect, we believe our thesis contributes to the study of natural language in financial disclosures and its effect on financial markets.

Chapter 3

Data, Data Treatment, and Variables

In this chapter we explain the data sources and preprocessing steps, define the variables used in the analysis, and show the most important descriptive statistics.

3.1 Data Sources and Data Preparation

We download daily financial time series data using Center for Research in Security Prices (CRSP) through Wharton Research Data Services (WRDS). Daily factor returns (Fama and French, 1993) together with daily risk-free return is downloaded from the Kenneth R. French data library[1]. We perform necessary cleaning of the financial time series[2] and correct for the appropriate adjustment factors that account for dividends and stock splits.

Of all the available types of corporate filings, we choose to analyze 10-K filings (annual reports). These reports are the only ones which are audited externally. Previous textual analysis research is indecisive of which filings are most informative. Loughran and McDonald (2011) use 10-Ks because they, on average, contain much more text and thus should be more suitable for textual analysis, while Jiang et al. (2019) include 10-Qs and argue that the increased frequency of data points is beneficial.

We download all SEC 10-K filings between January 1994 and December 2018 from the Notre Dame Software Repository For Accounting and Finance (http://sraf.nd.edu) created by Professor Bill McDonald, and first presented in Loughran and McDonald (2016). The 10-K filings are originally scraped from SEC's EDGAR database and have been parsed to exclude markup tags, ASCII-encoded graphics, and tables, as described more thoroughly on Professor McDonald's website[3]. Using the following criteria, we construct our data

[1] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[2] This includes replacing values of -99 (indicating unavailable volume data) in trading volume with ”N/A”, replacing price values of 0 (indicating unavailable price data) with ”N/A” and taking absolute values of all prices (negative values indicate that the prices are calculated as an average between the bid and ask price)

[3] https://sraf.nd.edu/data/stage-one-10-x-parse-data)

sample of 10-Ks:

- We only include 10-K filings from companies that have been listed on the S&P500 index any time during the period from 1994 and 2018. To do so, we use the *i0003* (S&P500 Comp-Ltd) Compustat - Capital IQ list through WRDS. We download company names, tickers, Central Index Keys (CIK), and Committee on Uniform Securities Identification Procedures (CUSIP) [4].
- To link financial time series data with SEC filings, we match using CUSIPs and CIKs. All filings which we are not able to match are removed from our sample.

The total number of unique firms amounts to 1,098, and the final sample consists of 15,700 10-K filings.

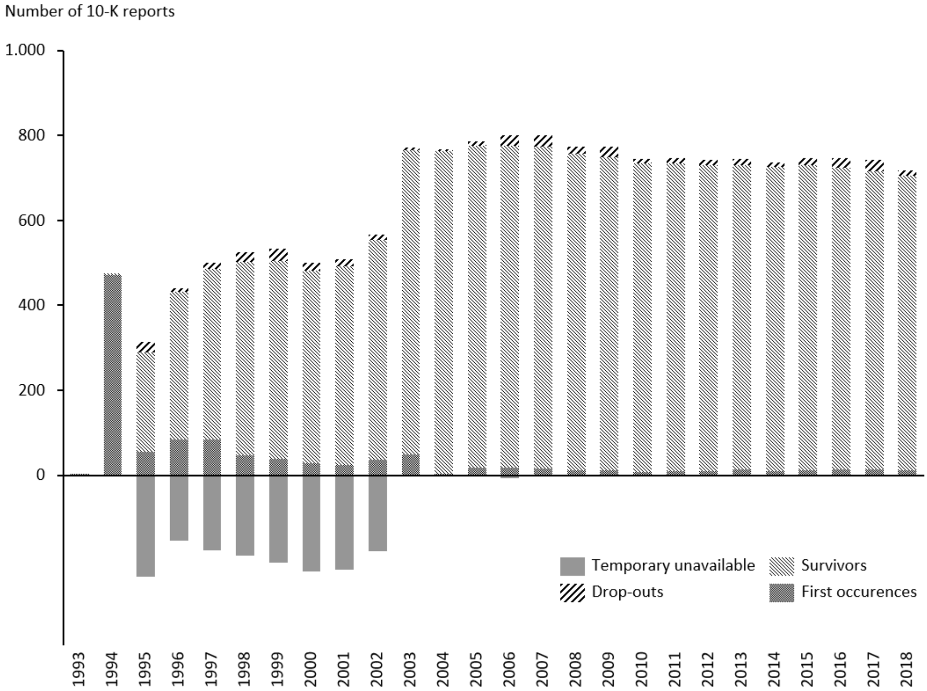


Figure 3.1: Number of reports available per year. Height of bar above x-axis represents the number of reports. Height of bar below the x-axis represents the number of reports that are unavailable because they are temporarily missing from the sample.

We choose to include the whole dataset in our analysis. From Figure 3.1 we can see that a large portion of the reports that first appeared in 1994 are missing from 1995 to 2002,

[4] CUSIPs and CIKs are both unique codes used to identify companies on different platforms. In contrast to tickers, these codes remain unchanged in the case of a company changing its name.

before re-appearing in 2003. Since, to our knowledge, every company is required to file a 10-K report annually, we consider the risk of selection bias to be small. It is more likely that they are missing due to some random technicality than some common, systematic attribute with the companies themselves. Nevertheless, we include the most important regressions for the sub-sample from 2004 and onwards in the appendix' Chapter A2. The essential conclusions in terms of signs and significance remain similar.

Important information that may affect company valuation should be contained in statements describing the future, so-called Forward-Looking Statements (FLS), and we, therefore, focus our analysis on these sentences. We also choose to use the whole report, as opposed to only the MD&As. Li (2010) analyzes FLS in MD&As and argues that one of the objectives in mandating MD&A sections is to provide meaningful and predictable information about future performances. Hence, investors and readers should receive truly descriptive information about trends that may affect forthcoming events. MD&As informational value are however questioned by Hüfner (2007) who criticizes MD&A regulations and its ability to inform investors about relevant future trends. Furthermore, a large number of companies add boilerplate (standarized text) sentences and generic language in their MD&As (SEC, 2003). Loughran and McDonald (2011) show empirically that analyzing the whole report gives better results than analyzing only MD&As.

Following the main steps from Li (2010), we define forward-looking sentences as follows:

1. We include all sentences that contain ”will”, ”should”, ”can”, ”could”, ”may”, ”might”, ”expect”, ”anticipate”, ”believe”, ”plan” , ”hope”, ”intend”, ”seek”, ”project”, ”forecast”, ”objective” or ”goal”. (The word ”shall”, which intuitively would be appropriate to include, is excluded from this list as it is used frequently in legal language and boilerplate disclosures.)

2. We exclude all sentences that meet one or more of the following criteria:

   (a) Consists only of majuscules (all capital letters)

   (b) Contains less than 5 words

   (c) The first character is either minuscule or not a letter

   (d) Contain ”undersigned”, ”herein”, ”hereinafter”, ”hereof”, ”hereon”, ”hereto”, ”theretofore”, ”therein”, ”thereof” and ”thereon” as these words are typically used in boilerplate language

   (e) Contains ”Expected”, ”anticipated”, ”forecasted”, ”projected” or ”believed” following after ”was”, ”were”, ”had” and ”had been”, as these sentences most likely are not forward-looking

   (f) Consists of more than 15% numerical characters

As discussed in Li (2010), the likelihood of not including sentences that are truly forward-looking is small (type-II error) because of the long list of search words used in the selection process. We will, on the other hand, most likely include some sentences which are, in fact, not forward-looking (type-I error).

3.2 Definition of Variables

3.2.1 Sentiment

We create a novel measure to quantify the sentiment level in a financial disclosure. We base it on the VADER model, but after modifying it to also include the dictionary from Loughran and McDonald (2011). In case the latter dictionary includes words that are already included in the VADER dictionary, we give priority to words from Loughran and McDonald (2011) by overwriting the word value in the original VADER dictionary. We assign all positive and negative words in Loughran and McDonald (2011) the maximum and minimum valence scores of $+4$ and -4, respectively. Our intention is to combine the comprehensive and advanced framework of VADER with the most well-documented dictionary for financial documents. This is in contrast to any of the conventional dictionary methods proposed so far (Tetlock et al., 2008; Kothari et al., 2009; Loughran and McDonald, 2011; Chen et al., 2013) which are based on the (proportional or TD-IDF-weighted) shares of positive and negative words in the texts.

The modified VADER outputs four sentiment variables: negative, neutral, positive, and compound. The first three scores are defined as ratios for the proportion of text falling into the respective categories. Hence they will always sum to 1. Of the three, we only use the positive and negative measures, which we denote as $vader\_pos$ and $vader\_neg$, respectively. In our analysis, we also include the compound measure, which returns a value ranging between -1 (most negative) to 1 (most positive). As valence scores range between -4 and 4, VADER uses a normalization formula to obtain the compound sentiment score of each sentence. Then we define the compound sentiment of an article i, $vader\_comp_i$, to be the average of normalized sentence compound scores.

For the sake of completeness and comparison, we also include two measures based on conventional word counts of the positive and negative words (denoted as $\{positive\}$ and $\{negative\}$, respectively) of the Loughran and McDonald (2011) dictionary. From previous literature (Kearney and Liu, 2014) we have chosen two of the most common measures (e.g. used by Twedt and Rees (2012)), which we call $naive\_tone$ and $tfidf\_tone$. The first is defined as follows:

$$naive\_tone_i = \frac{\sum_w \mathbb{I}(w \in \{positive\}) - \sum_w \mathbb{I}(w \in \{negative\})}{\sum_w \mathbb{I}(w \in \{positive\} \cup \{negative\})} \tag{3.1}$$

where $\mathbb{I}(\cdot)$ is an indicator function equal to 1 if the truth statement inside evaluates to true, and 0 otherwise, and $\sum_w$ means to sum over each word w in text i.

The second is defined similarly, with the exception that each term is weighted by their Term Frequency - Inverse Document Frequency (TF-IDF) values. TF-IDF is a common technique in NLP. A set of documents is represented by a matrix where each row corresponds to a document, and each column corresponds to a term. The elements in the matrix represent each words' weight in each document. There are several slightly different variations of how these weights are calculated, but the general structure is the same. ”TF” stands for ”term frequency”, and represents how often the term occurs in the document. Instead of using raw word counts, we follow Loughran and McDonald (2011) and use sub-linear (logarithmic) term frequencies. The intuition for this is that words frequently occurring in a document should have an impact that is less than proportional to the frequency, i.e., that ten occurrences of the word ”bad” is not ten times worse than one occurrence of the word ”sadly”. ”IDF” represents the inverse document frequency, i.e., that words commonly used across reports should be less important (and thus are down-weighted) relative to words more rarely used. The TF-IDF is the product of the ”TF” and ”IDF” scores of each word in each document, and we use these as weights on the words in Equation 3.1, resulting in the following formal definition of the measure we call $tfidf\_tone$:

$$tfidf\_tone_i = \frac{\sum_w W_{w,i}\mathbb{I}(w \in \{positive\}) - \sum_w W_{w,i}\mathbb{I}(w \in \{negative\})}{\sum_w W_{w,i}\mathbb{I}(w \in \{positive\} \cup \{negative\})} \tag{3.2}$$

where $W_{w,i}$ is defined as:

$$W_{w,i} = \begin{cases} (1 + ln(tf_i(w)) \cdot \left(1 + ln\frac{N}{df(w)}\right), & \text{if } tf_i(w) \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

and $tf_i(w)$ is the number of occurrences of word w in document i and $df(w)$ is the number of documents containing the word w. Following Wang et al. (2013), the weights are recalculated each year, as the document frequency of a specific word may vary across different years.

Loughran and McDonald (2011), find that using TF-IDF weights yield slightly more significant results relative to proportional weights (as in $naive\_tone$).

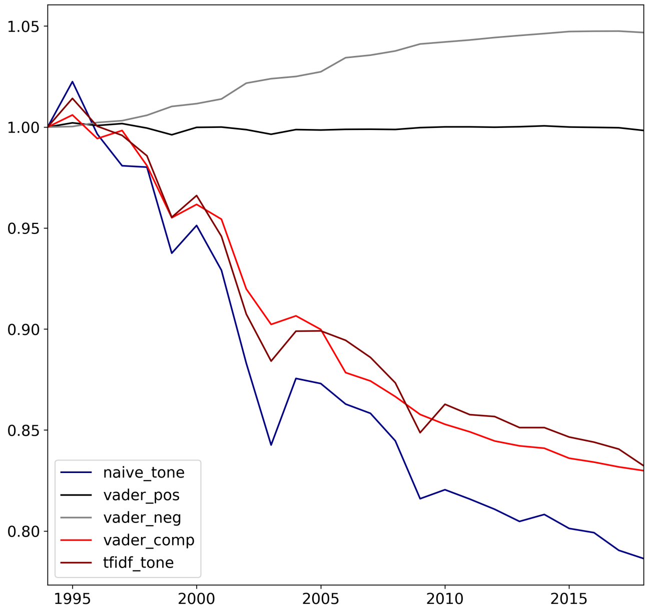


Figure 3.2: Average values for all sentiment measures, plotted over time. Each variable is normalized to start in 1. Decreasing lines means increasingly pessimistic sentiment with time, with the exception of $vader\_neg$ which has the opposite interpretation. The graph illustrates that it may be helpful to standardize the measures in order to better capture the cross-sectional differences.

We apply a differencing scheme where we subtract the arithmetic mean of the five previously published reports for the same company:

$$\Delta X_{i,t} = X_{i,t} - \mu(\{X_{i,t}\}_{t-6}^{t-1}) \tag{3.4}$$

where $X_{i,t}$ is either of the sentiment variables from company i on reporting day t. As Figure 3.2 shows, the sentiment measures exhibits clear signs of non-stationarity. Literature is not clear as to whether the levels themselves or changes in the levels relative to some historical reference should be used. Using some form of differencing could reduce the impact of contextually misclassified words, but at a possible cost of increasing the random variation in the frequency of common words (Loughran and McDonald, 2011). Since our time series is quite long and we are mainly interested in cross-sectional differences, we choose to apply differencing.

3.2.2 Readability

We consider three different measures for readability; the Fog index, the natural logarithm of the file-size of the report, and the number of Forward-Looking Statements (FLSs). Since they all intend to measures the same attribute, there should be no reason to assume that they are uncorrelated; hence we analyze them separately.

The Gunning Fog index aims to classify texts' readability into the number of required formal education years a person needs to comprehend the text. For example, a fog index of 17, indicates that a person needs to be a college graduate in order to have the appropriate reading level. A similar approach for interpreting the measure, e.g. used by Li (2008), is to classify the measures into five classes: Fog greater than 18 indicates that the text is unreadable; 14-18 (difficult); 12-14 (ideal); 10-12 (acceptable) and 8-10 (childish). The measure is calculated using the following formula:

$$gunning\_fog = 0.4 * \left[\frac{\#words}{\#sentences} + 100 * \left(\frac{\#complex\ words}{\#words} \right) \right] \tag{3.5}$$

where words with more than two syllabuses are defined as complex.

We define $log\_filesize$ as the natural logarithm of the number of ASCII bytes required to represent the parsed 10-K report from the EDGAR database, after XML tags and embedded binary data have been removed.

We also include the number of forward-looking sentences, num_sents, as a measure. This is somewhat similar to Karapandza (2016), who uses a measure based on the frequency of future tense words in reports and finds that firms talking less about the future generate positive abnormal returns of about 5% annually.

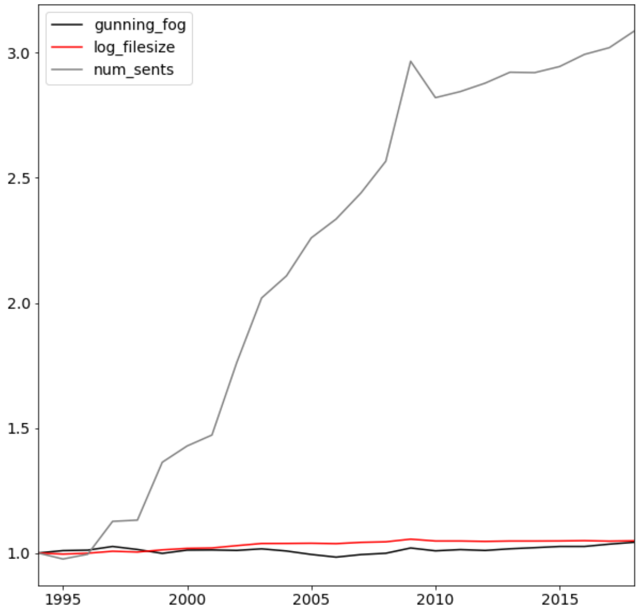


Figure 3.3: Average readability values per year. Variables are normalized to 1 in 1994. Increasing lines mean that reports get harder to read (have lower readability) with time.

Likewise as for the sentiment measures, we modify num_sents by subtracting the arithmetic mean of the five previously published reports for the same company:

$$\Delta num\_sents_{i,t} = num\_sents_{i,t} - \mu(\{num\_sents_{i,t}\}_{t-6}^{t-1}) \tag{3.6}$$

for company i on reporting day t. Dyer et al. (2017) report that investors have criticized financial disclosures for becoming more comprehensive and harder to read over the years. As Figure 3.3 shows, the num_sents variable is almost monotonically increasing. We do not modify $gunning\_fog$ and $log\_filesize$, as these appear more or less stationary.

Table 3.1 summarizes and describes the sentiment and readability variables that we use in the analysis.

Category	Variable	Differenced	Description
Sentiment			
	$\Delta naive\_tone$	Yes	Measure based on conventional word counts of positive and negative words of the Loughran and McDonald (2011) dictionary
	$\Delta tfidf\_tone$	Yes	Measure defined similarly as *naive_tone*, but each term is weighted by their TF-IDF values
	$\Delta vader\_pos$	Yes	Positive sentiment measure based on the VADER model, but modified to include the dictionary from Loughran and McDonald (2011)
	$\Delta vader\_neg$	Yes	Negative sentiment measure based on the VADER model, but modified to include the dictionary from Loughran and McDonald (2011)
	$\Delta vader\_comp$	Yes	Compound sentiment measure based on the VADER model, but modified to include the dictionary from Loughran and McDonald (2011)
Readability			
	Δnum_sents	Yes	Number of forward-looking sentences
	gunning_fog	No	Number representing the required number of formal education years a person needs to comprehend the text
	log_filesize	No	The natural logarithm of the number of ASCII bytes required to represent the parsed 10-K report

Table 3.1: Description of sentiment and readability variables

3.2.3 Abnormal Return

To capture the effect of stock price movements we first define the return of company i of day t as:

$$R_{i,t} = \frac{C_{i,t}}{C_{i,t-1}} - 1 \tag{3.7}$$

where $C_{i,t}$ is defined as the closing price for the stock of company i on day t.

To make sure that the abnormal returns are not driven by known sources of risk, we use a Fama-French 3-factor model from (Fama and French, 1993) to estimate the factor loadings $\beta_{MKT,i,t}$, $\beta_{HML,i,t}$ and $\beta_{SMB,i,t}$ from a rolling two-year (500 trading days) regression:

$$\begin{aligned} R_{i,t} - r_{f,t} =& \beta_{MKT,i,t} \cdot (r_{MKT,t} - r_{f,t}) \\ &+ \beta_{HML,i,t} \cdot r_{HML,t} + \beta_{SMB,i,t} \cdot r_{SMB,t} \quad t \in [-500; -1] \end{aligned} \tag{3.8}$$

The 1-day $\alpha_{i,t}$ for company i on day t is then calculated as the difference between the realized 1-day return and the linear prediction from the Fama and French (1993) model:

$$\begin{aligned} \alpha_{i,t} =& R_{i,t} - r_{f,t} - \beta_{MKT,i,t} \cdot (r_{MKT,t} - r_{f,t}) \\ &- \beta_{HML,i,t} \cdot r_{HML,t} - \beta_{SMB,i,t} \cdot r_{SMB,t} \end{aligned} \tag{3.9}$$

where $r_{mkt,t}$ is the return of the market portfolio at day t, $r_{HML,t}$ is the return of the high-minus-low-portfolio at day t, and $r_{SMB,t}$ is the return of the small-minus-big-portfolio on day t and $r_{f,t}$ is the risk-free return defined as the simple T-bill daily rate that, over the number of trading days in the month, compounds to the 1-month T-bill rate from Ibbotson and Associates Inc.[5]

3.2.4 Abnormal Volatility

We first estimate daily volatility using a German-Klass volatility estimator for the volatility of company i at day t, as discussed in Molnár (2012):

$$\widehat{\sigma^2_{GK,i,t}} = 0.5\left(ln\left(\frac{H_{i,t}}{L_{i,t}}\right)\right)^2 - (2ln2-1)\left(ln\left(\frac{C_{i,t}}{O_{i,t}}\right)\right)^2 + \left(ln\left(\frac{O_{i,t}}{C_{i,t-1}}\right)\right)^2 \tag{3.10}$$

where O, C, H, L is defined as the (appropriately adjusted) open price, closing price, high price, and low price, respectively.

Figure 3.4 shows a plot of the (annualized) volatility for the days around the filing date. It is evident that in general, volatility around the filing date does increase slightly on average, but only for the first few days.

[5] We get this data from the website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

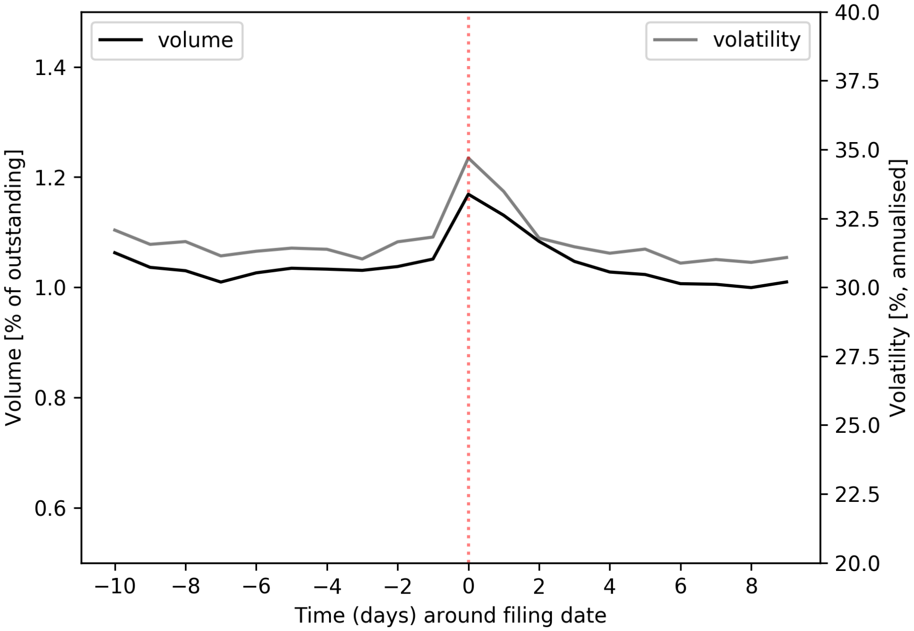


Figure 3.4: Volume and volatility around 10-K filing date. Volume (left) is represented as the number of traded shares as a percentage of the number of outstanding shares. Volatility (right) is defined as the square root of the 1-day German-Klass volatility from Equation 3.10, multiplied by $\sqrt{252}$ (annualized). Graph shows that the effect of the annual report only lasts a few days after the reports are publised.

To create a measure for the abnormal volatility, we establish the normal volatility as a one-month (22 trading days) trailing median of the natural log of the volatilities preceding a report, and then define abnormal volatility, $volatility_{i,t+d}$, as the difference between the logarithm of the observed volatility on day $t + d$ and the normal volatility established at day t. This way, we avoid polluting the normal volatility level with any potential influence from the report being published. At the same time, this step-wise sliding window reduces the validity for large values of d. As d becomes large, the time between the measured volatility and its reference period gets too large to be useful. This should not be a problem, however, due to the short-lived nature of the post-filing volatility observed in Figure 3.4.

3.2.5 Abnormal Volume

Figure 3.4 shows the average daily trading volume around filing dates as a percentage of outstanding shares. It exhibits a similar pattern as the volatility, with a sudden bump on the filing date, before gradually returning to a normal trading volume within the following five days.

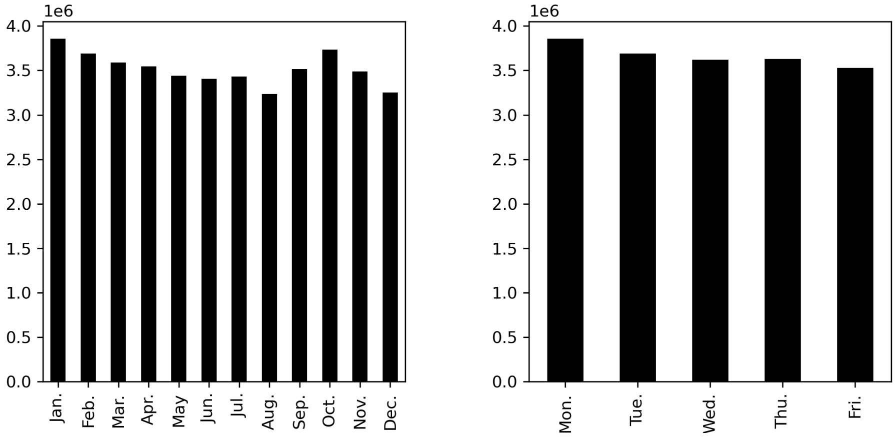


Figure 3.5: Average trading volume (number of shares traded) per week and month. It is clear that the underlying data must be adjusted for seasonalities.

From Figure 3.5, it is clear that there are seasonal variations in trading volume, both between weekdays and between months. We therefore first adjust the trading volume for seasonalities by calculating appropriate adjustment factors for each day of the week and month of the year for each company. The seasonality adjusted volume is then the product of the observed trading volume and the two relevant factors.

Similarly, as for abnormal volatility, we define abnormal volume as the difference between the normal daily trading volume, $NVLM_{i,t}$, of company i on trading day t, and the trading volume observed on that particular day. We define $NVLM_{i,t}$ to be the 22-day trailing median of the seasonality adjusted volume, $VLM_{i,t}$, prior to a report being published:

We then define the abnormal volume $volume_{i,t+d}$ for company i at day $t+d$ as the log difference between $VLM_{i,t+d}$ and $NVLM_{i,t}$.

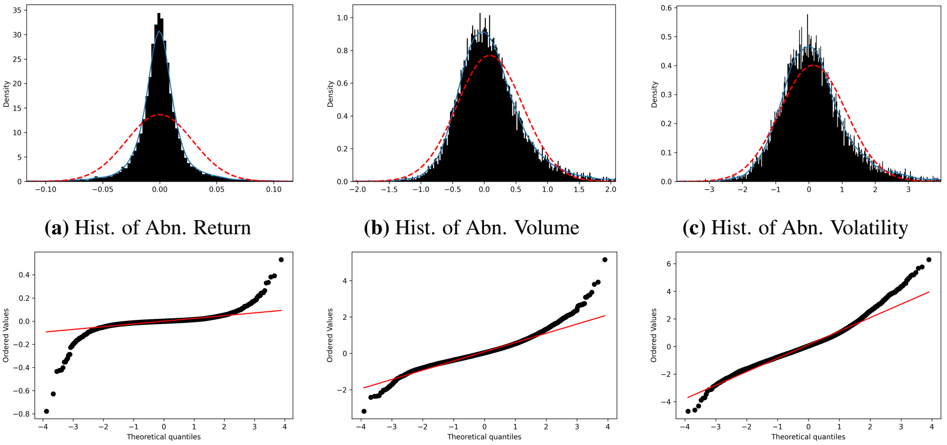
$$volume_{i,t+d} = \ln(VLM_{i,t+d}) - \ln(NLVM_{i,t}) \tag{3.11}$$

We choose this method to account for individual seasonality patterns in each company. The relatively short window for the trailing median is chosen to prevent the results from being contaminated by the non-stationary characteristics of trading volume for longer time spans. Similarly, as for volatility, this abnormal volume is only valid for reasonably small values of d.

3.3 Descriptive Statistics of Variables

Figure 3.6 and Table 3.2 shows the characteristics of the dependent variables for a selection of choices for d. We see that all variables are slightly leptokurtic.

3.3.1 Dependent Variables



(a) Hist. of Abn. Return **(b)** Hist. of Abn. Volume **(c)** Hist. of Abn. Volatility

(d) Q-Q plot of Abn. Return **(e)** Q-Q-plot of Abn. Volume **(f)** Q-Q plot of Abn. Volatility

Figure 3.6: Histogram and QQ-plots of dependent variables. Theoretical normal dist.(red-dotted line)/ Kernel Density(Blue line)

	count	mean	std	min	25%	50%	75%	max	skew	kurt
$return\_0$	13407	-0.000	0.029	-0.777	-0.009	-0.000	0.008	0.532	-2.482	96.768
$return\_3$	13410	-0.000	0.053	-0.941	-0.018	-0.000	0.017	1.712	3.111	148.897
$return\_30$	13378	0.001	0.110	-1.652	-0.047	0.002	0.050	0.906	-0.485	12.311
$return\_60$	13350	0.001	0.159	-2.041	-0.071	0.004	0.076	1.636	-0.444	11.228
$return\_180$	13183	-0.028	0.308	-5.267	-0.147	-0.013	0.112	6.021	-1.035	42.744
$return\_252$	13088	-0.038	0.391	-18.514	-0.179	-0.020	0.131	7.484	-8.433	404.264
$volume\_0$	14351	0.089	0.519	-3.175	-0.241	0.035	0.352	5.162	0.825	3.262
$volume\_3$	14345	0.053	0.500	-2.791	-0.248	0.027	0.326	3.634	0.475	3.206
$volume\_5$	14341	0.040	0.518	-16.726	-0.263	0.017	0.310	3.959	-1.847	79.216
$volume\_10$	14333	0.030	0.511	-2.866	-0.280	0.003	0.309	4.660	0.514	3.315
$volatility\_0$	14335	0.126	0.995	-4.688	-0.516	0.038	0.633	6.302	0.720	1.827
$volatility\_3$	14327	-0.003	0.875	-5.522	-0.563	-0.018	0.534	8.686	0.284	2.715
$volatility\_5$	14321	-0.008	0.869	-6.532	-0.570	-0.028	0.538	5.265	0.216	1.683
$volatility\_10$	14313	-0.035	0.893	-6.197	-0.605	-0.054	0.513	6.735	0.181	1.911

Table 3.2: Selected descriptive statistics for financial time series (10K). The table shows the dependent variables for a selection of choices for d (name ends with ”$\_d$”).

3.3.2 Independent Variables

Table 3.3 shows the descriptive statistics for our independent variables, and Table 3.4 shows the correlation matrix between the independent variables. The signs and sizes are as one would expect; negativity and positivity measures has correlations < 0, and measures for the same property (e.g. $\Delta naive\_tone$ and $\Delta vader\_comp$) has high correlations. Interestingly, $log\_filesize$ correlates 53% with $gunning\_fog$, implying that longer reports tend to also have a more complicated language.

	count	mean	std	min	25%	50%	75%	max	skew	kurt
$\Delta naive\_tone$	10697	-0.028	0.103	-0.787	-0.073	-0.020	0.028	0.625	-0.788	4.999
$\Delta tfidf\_tone$	10697	-0.022	0.102	-0.767	-0.067	-0.015	0.033	0.725	-0.747	5.171
$\Delta vader\_pos$	10697	-0.000	0.009	-0.095	-0.005	-0.000	0.005	0.083	-0.195	5.169
$\Delta vader\_neg$	10697	0.006	0.014	-0.078	-0.002	0.005	0.014	0.109	0.414	2.555
$\Delta vader\_comp$	10697	-0.023	0.065	-1.003	-0.058	-0.021	0.015	0.452	-0.989	13.727
Δnum_sents	10697	54.291	191.330	-1010.000	-31.400	33.800	117.400	3421.600	2.861	34.760
$gunning\_fog$	15700	21.871	2.113	12.700	20.549	21.576	22.855	51.720	1.680	11.661
$log\_filesize$	15700	12.952	0.611	8.504	12.575	12.947	13.314	16.233	0.021	1.188

Table 3.3: Readability and sentiment statistics

	$\Delta naive\_tone$	$\Delta tfidf\_tone$	$\Delta vader\_pos$	$\Delta vader\_neg$	$\Delta vader\_comp$	Δnum_sents	gunning_fog	log_filesize
$\Delta naive\_tone$		**0.838**	**0.433**	**-0.487**	**0.571**	-0.068	-0.105	0.041
$\Delta tfidf\_tone$	**0.772**		**0.354**	**-0.353**	**0.405**	-0.188	-0.182	-0.051
$\Delta vader\_pos$	**0.388**	**0.295**		**-0.205**	**0.619**	0.117	0.030	0.124
$\Delta vader\_neg$	**-0.498**	**-0.349**	**-0.205**		**-0.833**	-0.121	-0.150	**-0.208**
$\Delta vader\_comp$	**0.569**	**0.383**	**0.591**	**-0.832**		0.181	0.122	**0.248**
Δnum_sents	-0.102	**-0.261**	0.096	-0.109	0.163		**0.265**	**0.460**
$gunning\_fog$	-0.045	-0.133	0.062	-0.189	**0.204**	**0.245**		**0.533**
$log\_filesize$	0.009	-0.091	0.080	**-0.210**	**0.219**	**0.387**	**0.578**	

Table 3.4: Readability and sentiment correlation matrix (Spearman above/Pearson below). Correlations larger than 20% are in bold.

Chapter 4

The Impact of Sentiment and Readability

4.1 Method

We specify a simple, pooled, linear regression model to capture the effect from the sentiment and readability of 10-K filings.

$$Y_{i,t} = \beta\overline{X_{i,t}} + \{year_y\}_{y=1993}^{2018} + \epsilon \tag{4.1}$$

The dependent variable $Y_{i,t}$ can be either $return_{i,t}$, $volatility_{i,t}$ or $volume_{i,t}$. The independent variable X takes the value from one of our four sentiment measures ($\Delta naive\_tone$, $\Delta vader\_comp$, $\Delta vader\_pos$ or $\Delta vader\_neg$) or three readability measures ($gunning\_fog$, Δnum_sents and $log\_filesize$) from Table 3.1. Since the variables are correlated, the regressions must be performed separately. However, in each case, we normalize the independent variables:

$$\overline{X_{i,t}} = \frac{X_{i,t} - \mu(X)}{\sigma(X)} \tag{4.2}$$

This allows for comparison between the coefficients, as they can be interpreted as having the unit of ”impact per standard deviation from the mean”. To further isolate the effect of just the inter-company differences in filings, we add dummy variables for each year. Although there should be no theoretical reason mandating this, we observe that these dummy variables significantly change the results. For completeness, we include the results for return without Year-dummies in Table A2.4.

We carry out the regressions for a selection of trading days after the reports are published. For abnormal return, we include up to 252 trading days (i.e., about one calendar year), and report the abnormal return on a cumulative basis from (and including) the filing date to the given day. Abnormal volatility and volume are reported noncumulatively for the first ten trading days only, as Figure 3.4 on page 21 indicate that all abnormal movements happen well within this timeframe.

We use standard OLS, but with robust (heteroscedasticity-consistent (HC)) standard errors of the type HC3 from MacKinnon and White (1985)[1]

4.2 Results

4.2.1 Abnormal Returns

Table 4.1 shows the compiled results for how sentiment and readability affect abnormal returns for a selection of days after the reports are published. Coefficients are reported as the accumulated percentage period abnormal return per standard deviation from the mean measure value.

Compared with the average sentiment in reports, we observe that one-year abnormal return decreases 1.34% per standard deviation of negativity ($\Delta vader\_neg$). However, the degree of positivity of a report ($\Delta vader\_pos$) seems to have no significant effect on abnormal return. After taking a conventional word-counting approach, Loughran and McDonald (2011) and others attribute the lack of incremental value in the positive word list to the fact that negative news is often framed as negations of something positive (e.g., "did not benefit"), while the contrary is not equally common (e.g., "not downgraded" is rarely used). These negations are taken care of in the VADER algorithm, but the results remain the same. Therefore, we think a more likely explanation is that negative news is often padded with positive words to "soften the blow".

Consequently, the combined metric of negativity and positivity ($\Delta vader\_comp$) yields results that are less significant than the level of negativity on its own. It is also interesting to note that there is a considerable delay of up to a year from the time a report is published to the abnormal returns become significantly non-zero. This could indicate that the reports seem to contain important sentiment information that investors fail to take into account immediately after publication. This delay is in contrast to what Loughran and McDonald (2011) find by using a different sample of companies over a different period and defining abnormal returns slightly differently. They find significant relationships in the short term but cannot obtain statistically significant relationships for 1-year returns.

Our modified VADER measure does seem to be more efficient than both $\Delta naive\_tone$ and $\Delta tfidf\_tone$ in capturing the variations of sentiment that give rise to the differences in abnormal return. Also, contrary to Loughran and McDonald (2011), we find that the use of TF-IDF weights yields less significant results[2].

Readability also seems to have a positive long-term effect on abnormal return. At 180 and 252 trading days after the reports are published, all readability measures yield significant, negative coefficients, indicating that a longer and less clear 10-K-report is

[1] Robust (heteroscedasticity-consistent (HC)) standard errors of the type HC3 from MacKinnon and White (1985) is defined as:

$$(X^T X)^{-1} X^T diag(e_i^2/(1-h_{ii})^2) X (X^T X)^{-1} \tag{4.3}$$

where $h_{ii} = x_i(X^T X)^{-1} x_i^T$.

[2] Loughran and McDonald (2011) do not use time-standardized variables. When using the non-standardized version of the measure, we get similar results as they do.

associated with negative abnormal return the following year. In line with the findings of Loughran and McDonald (2014), the Fog index seems to be an inferior measure of readability compared to the log of the filesize, in terms of capturing variance in abnormal financial returns.

There are two likely explanations for the observation that readability and return are positively related. Li (2008) argues that firms may use longer reports to conceal damaging information strategically. Bloomfield (2008) further hypothesizes that poorly performing firms may need more text to explain the company situation more thoroughly.

An alternative explanation takes a risk-based view. Conventional finance theories (Easley and O'Hara, 2004) predict that ceteris paribus, less public information is riskier. Thus, investors must be compensated for this increased risk by receiving higher returns on these investments. When risk increases, the stock price must fall to accommodate the increased required rate of return, resulting in a negative abnormal return. If longer reports deliver *less* public information due to the decrease in information quality, this could explain why longer reports are negatively related to subsequent abnormal returns. This is also in line with the findings from Loughran and McDonald (2014).

Executives that want to maximize company valuations should decrease the risk of the companies. Our findings may suggest that one way of decreasing this risk is to write short and clear reports.

We should be careful to point out that we do not claim any causal relationship between sentiment or readability and abnormal return. It might also be that the sentiment and readability measures proxy for other contemporaneous information and that this is what drives the results. Furthermore, the adjusted R2 is very low, about 1% in all regressions.

	Dependent variable: $return_t$								
	(1 days)	(3 days)	(5 days)	(10 days)	(30 days)	(45 days)	(60 days)	(180 days)	(252 days)
$\Delta naive\_tone$	0.0541	0.0620	0.0349	-0.0093	0.0278	0.0162	0.0937	0.5693*	0.7613**
	(1.555)	(1.343)	(0.665)	(-0.140)	(0.250)	(0.114)	(0.552)	(1.792)	(1.968)
$\Delta tfidf\_tone$	0.0081	0.0353	0.0168	-0.0529	-0.0383	-0.1129	0.0055	0.4422	0.5765
	(0.235)	(0.774)	(0.321)	(-0.789)	(-0.340)	(-0.797)	(0.032)	(1.382)	(1.441)
$\Delta vader\_pos$	0.0382	0.0585	0.0174	0.0282	0.0633	0.1521	0.2615*	0.3118	0.6260
	(1.038)	(1.191)	(0.318)	(0.420)	(0.588)	(1.141)	(1.704)	(1.014)	(1.496)
$\Delta vader\_neg$	-0.0177	-0.0290	0.0166	-0.0002	-0.0472	0.0229	-0.2565	**-0.9176*****	**-1.3442*****
	(-0.403)	(-0.494)	(0.267)	(-0.003)	(-0.431)	(0.159)	(-1.532)	(-2.742)	(-3.354)
$\Delta vader\_comp$	0.0473	0.0482	-0.0015	0.0281	0.1035	0.0403	0.2930*	**0.8480*****	**1.2608*****
	(1.250)	(0.968)	(-0.027)	(0.407)	(0.960)	(0.284)	(1.797)	(2.666)	(3.179)
Δnum_sents	-0.0070	-0.0460	-0.0540	-0.0414	-0.0471	-0.0284	-0.0322	-0.6233**	-0.7235**
	(-0.144)	(-0.748)	(-0.763)	(-0.486)	(-0.346)	(-0.176)	(-0.184)	(-2.008)	(-1.968)
$gunning\_fog$	0.0001	-0.0072	0.0723	0.0757	-0.0553	0.0019	0.0350	**-0.7202*****	-0.7164*
	(0.003)	(-0.166)	(1.492)	(1.211)	(-0.566)	(0.016)	(0.256)	(-2.654)	(-1.949)
$log\_filesize$	-0.0612*	-0.0799	-0.0389	-0.0509	-0.1240	-0.0578	-0.0331	**-1.1092*****	**-1.5917*****
	(-1.793)	(-1.623)	(-0.740)	(-0.756)	(-1.187)	(-0.462)	(-0.229)	(-3.667)	(-3.692)
Observations	14381	13410	13407	13398	13378	13369	13350	13183	13088
Δ Observations	9847	9753	9750	9743	9721	9708	9694	9568	9488
Note:									*p<0.1; **p<0.05; ***p<0.01 *(t scores in parentheses)*

Table 4.1: Impact of sentiment and readability on abnormal return. Regression is specified as $return_{i,t,d} = \beta \overline{X_{i,t}} + \{year_y\}_{y=1993}^{2018} + \epsilon$, where $X_{i,t}$ takes any of the independent variables. The regressions are performed on each dependent variable individually. Table shows the beta coefficients with its respective t-scores. Each row and column are separate regressions, where the rows represent the independent variable and the columns represent the time horizon of the dependent variable. Δ Observations is number of observations for differenced variables (name starts with "Δ "), while Observations is number of observations for the other regressions.

4.2.2 Abnormal Volatility and Volume

As we show in Table 4.2, abnormal volatility increases with $gunning\_fog$ and $log\_filesize$ the first days after the reports are published, before returning to normal on day five. Note that $log\_filesize$, when calculating abnormal volatility, does not take into account differences in firm complexity and size. Intuitively, larger corporations require more words to describe the business and operations adequately, ceteris paribus. It is therefore fair to argue that one should control for firm size by including market cap as an explanatory variable together with $log\_filesize$. However, we do not find that the market cap, to a significant extent, explains the variation in file size, nor does it change the coefficients for $log\_filesize$. This may be because our dataset consists of S&P-500 companies, which, per definition, is a collection of the largest listed companies. We, therefore, choose not to keep market cap as an explanatory variable.

None of the sentiment measures seem to significantly influence the level of abnormal volatility.

	Dependent variable: volatility$_t$										
	(0 days)	(1 days)	(2 days)	(3 days)	(4 days)	(5 days)	(6 days)	(7 days)	(8 days)	(9 days)	(10 days)
$\Delta naive\_tone$	0.0990	0.6185	0.9885	-1.0592	0.6905	0.1394	0.4349	-0.7311	-0.7931	-0.3462	-1.7848**
	(0.103)	(0.671)	(1.151)	(-1.265)	(0.836)	(0.160)	(0.504)	(-0.859)	(-0.928)	(-0.384)	(-2.030)
$\Delta tfidf\_tone$	0.7540	0.1435	0.8203	-0.9279	0.3205	0.3507	-0.7709	-1.0755	-1.8911**	-1.5283*	-1.2842
	(0.803)	(0.160)	(0.964)	(-1.114)	(0.382)	(0.391)	(-0.907)	(-1.231)	(-2.256)	(-1.698)	(-1.451)
$\Delta vader\_pos$	-0.2741	-0.3080	0.9421	-0.2273	0.1730	0.1394	-0.2764	-0.4613	-0.3159	-1.7332**	-1.6780*
	(-0.279)	(-0.337)	(1.150)	(-0.284)	(0.212)	(0.164)	(-0.333)	(-0.543)	(-0.381)	(-2.014)	(-1.919)
$\Delta vader\_neg$	-0.3962	-0.0552	-0.5204	-0.5151	-0.5626	-0.3317	-1.0954	-0.3923	-0.2340	-0.1188	0.5657
	(-0.387)	(-0.057)	(-0.601)	(-0.617)	(-0.657)	(-0.371)	(-1.275)	(-0.456)	(-0.267)	(-0.136)	(0.643)
$\Delta vader\_comp$	0.1330	-0.0083	0.3547	-0.0221	0.3851	-0.2299	0.6476	-0.0493	0.3430	-0.5536	-1.5544*
	(0.134)	(-0.009)	(0.440)	(-0.027)	(0.463)	(-0.269)	(0.768)	(-0.059)	(0.402)	(-0.655)	(-1.812)
Δnum_sents	0.2947	1.6960*	1.0789	1.5391*	2.0172**	0.7554	0.5388	1.7699**	1.5265*	0.3140	1.9283**
	(0.287)	(1.797)	(1.167)	(1.899)	(2.492)	(0.960)	(0.696)	(2.215)	(1.765)	(0.402)	(2.281)
$gunning\_fog$	1.6998**	1.8671**	1.4819**	**2.2511*****	**2.1605*****	0.8682	-0.2033	0.2493	0.5142	-0.0184	1.5926**
	(2.123)	(2.422)	(2.012)	(3.117)	(2.936)	(1.173)	(-0.276)	(0.340)	(0.694)	(-0.024)	(2.090)
$log\_filesize$	**3.7099*****	**4.2238*****	**3.0368*****	**3.1152*****	**2.5330*****	1.2899*	0.4377	1.0292	1.5602**	-0.4469	1.2526
	(4.368)	(5.130)	(3.905)	(4.076)	(3.333)	(1.649)	(0.559)	(1.326)	(1.992)	(-0.568)	(1.584)
Observations	14335	14324	14328	14327	14324	14321	14316	14312	14315	14316	14313
Δ Observations	9824	9819	9819	9818	9816	9814	9811	9808	9810	9810	9808
Note:											*p<0.1; **p<0.05; ***p<0.01 *(t scores in parentheses)*

Table 4.2: Impact of sentiment and readability on abnormal volatility. Regression is specified as $volatility_{i,t} = \beta\overline{X_{i,t}} + \{year_y\}_{y=1993}^{2018} + \epsilon$, where $X_{i,t}$ takes any of the independent variables. The regressions are performed on each dependent variable individually. Table shows the beta coefficients with its respective t-scores. Each row and column are separate regressions, where the rows represent the independent variable and the columns represent the time horizon of the dependent variable. Δ Observations is number of observations for differenced variables (name starts with "Δ "), while Observations is number of observations for the other regressions.

We show the results for abnormal volume in Table 4.3. The coefficient of $log\_filesize$, the only properly significant measure, is positive throughout the ten days but is decreasing in size and significance throughout the period. This indicates that longer reports, and reports that have more future tense sentences, make investors disagree about how this extra information should be interpreted in terms of valuation, increasing the level of trading.

	Dependent variable: volume$_t$										
	(0 days)	(1 days)	(2 days)	(3 days)	(4 days)	(5 days)	(6 days)	(7 days)	(8 days)	(9 days)	(10 days)
$\Delta naive\_tone$	0.3244 (0.652)	0.3042 (0.608)	0.8534* (1.781)	0.6047 (1.289)	0.6472 (1.412)	0.6520 (1.394)	0.1130 (0.245)	0.4266 (0.920)	0.6241 (1.380)	0.6466 (1.339)	0.3305 (0.681)
$\Delta tfidf\_tone$	0.5020 (1.023)	0.3808 (0.750)	0.6345 (1.316)	0.4716 (1.016)	0.5078 (1.117)	0.3359 (0.725)	-0.2601 (-0.564)	0.2081 (0.439)	0.2767 (0.599)	-0.1895 (-0.387)	0.0707 (0.142)
$\Delta vader\_pos$	0.5099 (1.061)	-0.0169 (-0.035)	0.4318 (0.960)	0.3895 (0.884)	0.4234 (0.917)	0.2338 (0.496)	-0.1157 (-0.264)	0.5924 (1.321)	0.3480 (0.778)	-0.1491 (-0.327)	0.2306 (0.485)
$\Delta vader\_neg$	-0.2106 (-0.408)	-0.1847 (-0.358)	-0.5741 (-1.168)	-0.4848 (-1.040)	-0.2148 (-0.434)	0.0684 (0.142)	-0.2660 (-0.569)	-0.5548 (-1.171)	-0.5943 (-1.294)	-0.5989 (-1.244)	-0.6251 (-1.333)
$\Delta vader\_comp$	0.3708 (0.724)	0.0096 (0.020)	0.3857 (0.852)	0.6433 (1.452)	0.4563 (0.950)	0.1207 (0.263)	0.1923 (0.427)	0.6157 (1.350)	0.5044 (1.118)	0.3466 (0.735)	0.4655 (1.017)
Δnum_sents	0.4061 (0.835)	1.0129** (2.074)	0.9188** (2.017)	1.0043** (2.268)	1.0357** (2.057)	**1.2289***** (2.789)	0.5320 (1.231)	0.6114 (1.340)	0.6833 (1.457)	0.3501 (0.810)	0.4142 (0.894)
$gunning\_fog$	0.2029 (0.450)	0.5007 (1.078)	0.6962 (1.574)	0.6717 (1.574)	0.5965 (1.312)	-0.0166 (-0.035)	-0.4763 (-1.080)	-0.0641 (-0.150)	-0.0045 (-0.010)	0.4014 (0.864)	0.9422** (2.067)
$log\_filesize$	**1.6976***** (3.673)	**1.8420***** (3.792)	**1.8101***** (4.041)	**1.4733***** (3.237)	**1.7698***** (3.741)	1.1611** (2.486)	0.0427 (0.094)	1.0811** (2.345)	0.9440** (2.007)	0.4127 (0.864)	1.0353** (2.173)
Observations	14351	14347	14347	14345	14344	14341	14337	14336	14338	14335	14333
Δ Observations	9825	9822	9820	9819	9818	9815	9814	9812	9813	9811	9810
Note:											*p<0.1; **p<0.05; ***p<0.01 *(t scores in parentheses)*

Table 4.3: Impact of sentiment and readability on abnormal trading volume. Regression is specified as $volume_{i,t} = \beta\overline{X_{i,t}} + \{year_y\}_{y=1993}^{2018} + \epsilon$, where $X_{i,t}$ takes any of the independent variables. The regressions are performed on each dependent variable individually. Table shows the beta coefficients with its respective t-scores. Each row and column are separate regressions, where the rows represent the independent variable and the columns represent the time horizon of the dependent variable. Δ Observations is number of observations for differenced variables (name starts with "Δ "), while Observations is number of observations for the other regressions.

Chapter 5

The Impact of Topics from Latent Dirichlet Allocation

In this chapter, we present the method we use to extract topics from the SEC filings, along with a description of the necessary steps for preprocessing.

5.1 10-K Preprocessing

Textual data available in natural language form is not suitable for use with data mining techniques. Natural language contains words that are inflected (words expressed in distinct grammatical terms), so to produce representative insight from the text data, we need to use techniques and principles from text mining and NLP. Following Tripathi et al. (2015), we use tokenization, stop words filtering and lemmatization. Additionally, we create bigrams to capture two-word sequences. The following steps describe how we preprocess tweets:

1. **Tokenization:** Each SEC filing is divided into its fundamental structure (words).

2. **Make bigrams:** A bigram is an arrangement of two elements (words) that repeatedly occur together in a text document. ”common stock”, for example, makes the bigram ”common_stock”.

3. **Stop words filtering:** Filtering out ”stop words”, i.e., words that are too common in the English language to contribute to separating the SEC filings into different topics. The Python package `nltk.corpus` includes a list of English stopwords, which we use for this purpose.

4. **Lemmatization:** Lemmatization is a text normalization technique used to reduce the inflection in words and makes it possible to map a collection of words to the identical root. The technique groups together different inflected forms of a word and reduces the inflected words properly to ensure that the root word belongs to the language. The output of lemmatization is a proper word that benefits us when we

interpret the topics generated by our LDA algorithm. ”Gone”, ”going” and ”went” are, for instance, reduced to the root ”go”. For lemmatization, we use the Python package `spacy.lemmatizer`.

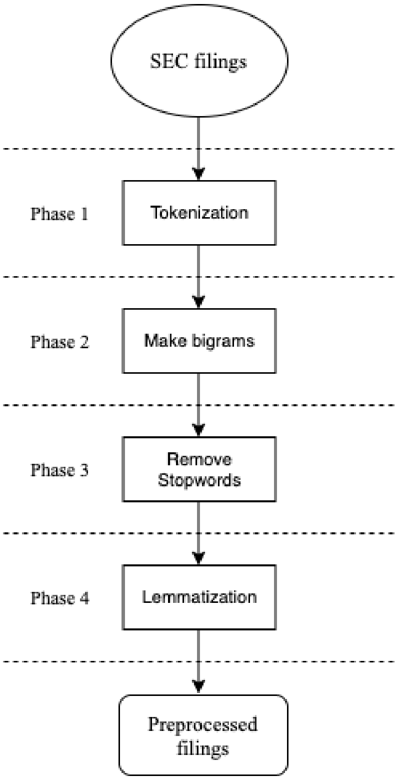


Figure 5.1: Illustration of 10-K text prepossessing

5.2 The LDA Method on 10-K filings

We use LDA to identify topics automatically and to infer the topic distribution of each SEC filing. In this subsection, we present the model from the original paper on LDA (Blei et al., 2003) in the context of classifying SEC filings, and we use the terminology adopted by the same paper.

Each word in the corpus is denoted as a V-dimensional indicator vector, with one and only one element equal to 1 and the rest 0, such that $w^v = 1$ for the vth word and $w^u = 0$ for $u \neq v$. We make the simplification of not considering bigrams in this explanation, but the concept is easily extended to handle this by treating each bigram as a separate word. A document is represented by an ordered list of N words $\vec{w} = (w_1, w_2, ..., w_N)$, and a corpus is a set of M documents denoted as $D = \{\vec{w}_1, \vec{w}_2, ..., \vec{w}_M\}$.

The objective is to create a probabilistic model of the probability distributions for topics in documents and words in topics. This is achieved by representing documents as random mixtures of latent topics, and where each topic is represented by a distribution of words. The algorithm assumes the following generative process for each $\vec{w} \in D$:

1. $Choose\ N \sim Poisson(\xi)$
2. $Choose\ \Theta \sim Dir(\alpha)$
3. For each of the N words $w_{n \in [1..N]}$:
   - Choose topic $z_n \sim$ Multinominal(Θ)
   - Choose word w_n from $p(w_n|z_n, \beta)$

β is here a $k \times V$ matrix where $\beta_{i,j} = p(w^j = 1|z^i = 1)$ (representing the weights of a word j in a topic i), and Θ is a k-dimensional Dirichlet random variable (representing the distribution of topics in a document) with the following property density:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \Theta_1^{\alpha_1 - 1} ... \Theta_k^{\alpha_{k-1}} \tag{5.1}$$

where α is a k-vector of strictly positive parameters α_i (representing the prior weight of a topic i in a document).

The joint distribution of a topic mixture Θ, the set of N topics $\vec{z}$ and the set of N words $\vec{w}$ is given by:

$$p(\theta, \vec{z}, \vec{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\Theta) p(w_n|z_n, \beta) \tag{5.2}$$

By integrating out Θ and summing over z, we obtain the marginal distribution of a document (we assume that the order of words is irrelevant, commonly known as the ”Bag-of-word”-assumption),

$$p(\vec{w}|\alpha, \beta) = \int p(\Theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\Theta) p(w_n|z_n, \beta) \right) d\Theta \tag{5.3}$$

The corpus probability is then the product of the marginal probabilities of the single documents (assuming that the order of the documents is irrelevant):

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\Theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\Theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\Theta_d \tag{5.4}$$

The parameters α and β are estimated on a corpus of documents (D) by maximizing the log-likelihood of the data using an iterative EM[1] algorithm:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \text{log} p(\vec{w}_d|\alpha, \beta) \tag{5.5}$$

[1] An EM algorithm is an iterative method that works by alternating between calculating the expectation (the E-step) and maximising (the M-step) the likelihood of the observed data with respect to the parameters.

The end goal is to retrieve the posterior distribution of the hidden variables (Θ and $\vec{z}$) given a document $\vec{w}$, $p(\Theta, \vec{z}|\vec{w}, \alpha, \beta)$. As exact inference is generally intractable, approximation techniques are used to get approximate estimates.

To overcome the issue of maximum likelihood assigning zero probability to unseen words (for out of sample classification), a ”smoothing technique” is used to assign a positive probability to those words.

When the model is built, it takes a pre-processed document (a company filing in our case) as input, and outputs the probability distribution describing the probability that the filing belongs to each of the topics. The process is illustrated in Figure 5.2 and 5.3.

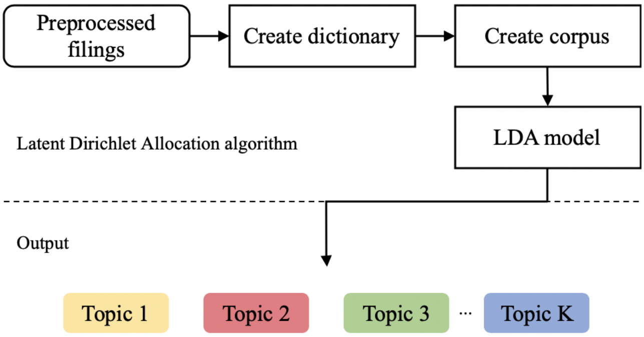
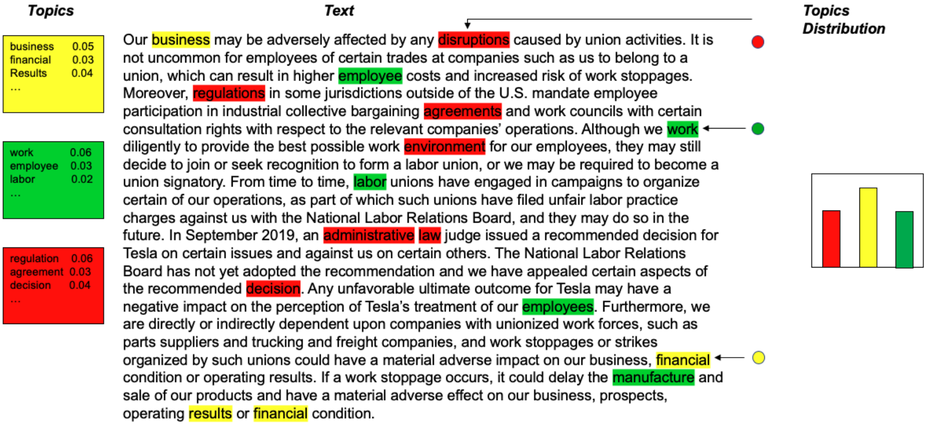


Figure 5.2: Illustration of LDA Algorithm



Our business may be adversely affected by any disruptions caused by union activities. It is not uncommon for employees of certain trades at companies such as us to belong to a union, which can result in higher employee costs and increased risk of work stoppages. Moreover, regulations in some jurisdictions outside of the U.S. mandate employee participation in industrial collective bargaining agreements and work councils with certain consultation rights with respect to the relevant companies' operations. Although we work diligently to provide the best possible work environment for our employees, they may still decide to join or seek recognition to form a labor union, or we may be required to become a union signatory. From time to time, labor unions have engaged in campaigns to organize certain of our operations, as part of which such unions have filed unfair labor practice charges against us with the National Labor Relations Board, and they may do so in the future. In September 2019, an administrative law judge issued a recommended decision for Tesla on certain issues and against us on certain others. The National Labor Relations Board has not yet adopted the recommendation and we have appealed certain aspects of the recommended decision. Any unfavorable ultimate outcome for Tesla may have a negative impact on the perception of Tesla's treatment of our employees. Furthermore, we are directly or indirectly dependent upon companies with unionized work forces, such as parts suppliers and trucking and freight companies, and work stoppages or strikes organized by such unions could have a material adverse impact on our business, financial condition or operating results. If a work stoppage occurs, it could delay the manufacture and sale of our products and have a material adverse effect on our business, prospects, operating results or financial condition.

Figure 5.3: Illustration of LDA Topic Distribution

5.3 Defining the Appropriate Number of Topics

The LDA model does not inform us what the number of topics of a corpus should be. A necessary input, and maybe the most important decision of building our LDA model is, therefore, defining the number of topics, K. One commonly used measure is ”perplexity” (e.g. used by Dyer et al. (2017) and Blei et al. (2003)), which is formally defined in Equation A3.1. However, the perplexity will, in most cases, continue to diminish with an increasing number of topics, so choosing the optimal number of topics will still require manual consideration of where the graph flattens out. Others suggest using topic coherence to identify the appropriate number of topics (Newman et al., 2010; Mimno et al., 2011), which measures the semantic correlation between the characteristic words. There are various approaches used to calculate the coherence score, and the different methods are carried out in detail in a study by Röder et al. (2015). Following Röder et al. (2015), we use the terminology adopted by the same paper, and select the coherence measure that yields the best performance (denoted C_V). The literature shows that coherence based on Normalized Pointwise Mutual Information (NPMI), a quantified measure of association used in statistics and an extension of Pointwise Mutual Information (PMI), gives the strongest correlation with human interpretation (Bouma, 2009; Röder et al., 2015; Aletras and Stevenson, 2013).

The LDA model generates K topics where each topic consists of top-n characteristic words. Furthermore, we let T denote a topic represented by n words ($T = w_1, w_2, ..., w_n$) and follow Aletras and Stevenson (2013) to calculate PMI and NPMI:

$$\text{PMI}(w_i, w_j) = \log\left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right) \qquad w_i, w_j \in T \tag{5.6}$$

$$\text{NPMI}(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log(P(w_i, w_j))} \qquad w_i, w_j \in T \tag{5.7}$$

An extrinsic corpus is used as a reference to recognize context characteristics and accumulate word frequencies, and the probabilities are based on the collected co-occurrence counts (Newman et al., 2010).

We choose to base our analysis on $K = 50$, as this is the number of topics that achieves the highest coherence score (0.514). To calculate the topic coherence, C_V, we use the Python package `gensim.models.CoherenceModel` and run our LDA model for different values of $K \in [15; 105]$. We calculate the score for each iteration (see Figure 5.4) and examine the results generated. For a selection of other K, we inspect the resulting topics manually to infer wether the resulting topics makes sense. The classes are to a large extent recognizable between different values of K, but in our opinion, $K = 50$ yields the most well organized topics.

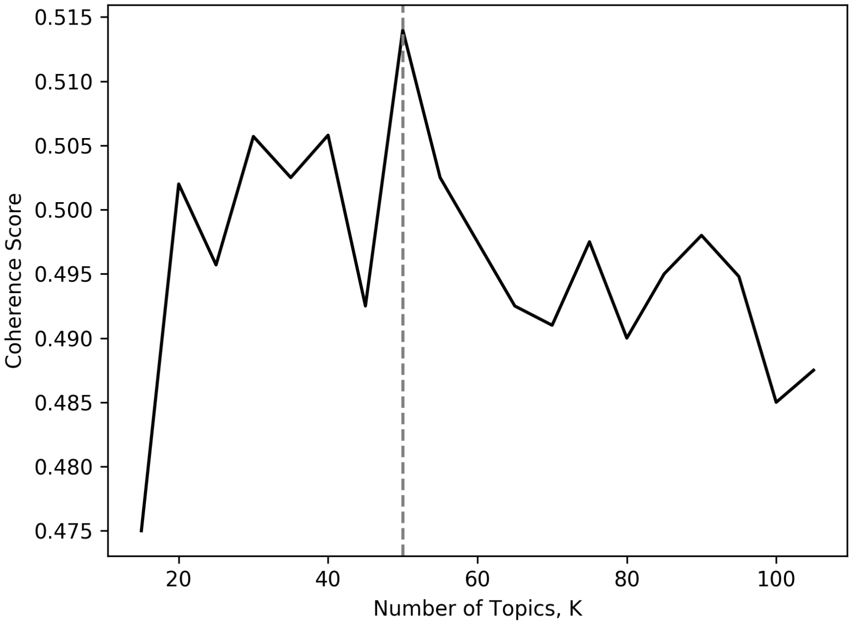


Figure 5.4: Coherence score for different number of topics. Higher value means more coherent topics. Dotted line shows $K = 50$, which is the number of topics where we achieve the maximum coherence score.

5.4 10-K Topic Names

The output of the LDA model is, as mentioned, distribution of topics with size K, each containing a set of characteristic words. These topics arise from statistical properties only and do not necessarily have anything semantically in common. Human interpretation is therefore needed to generate topic names. We summerize the resulting topics with names in Figure 5.5 and Figure 5.6 as 50 word clouds, where the relative size of each word represents the relative importance of that word in the given topic. We provide a complete list of the 50 topics and their associated keywords in A3.1. To infer the meaning of each topic, we also investigate the words in a typical context by inspecting the archetype sentences for each topic.

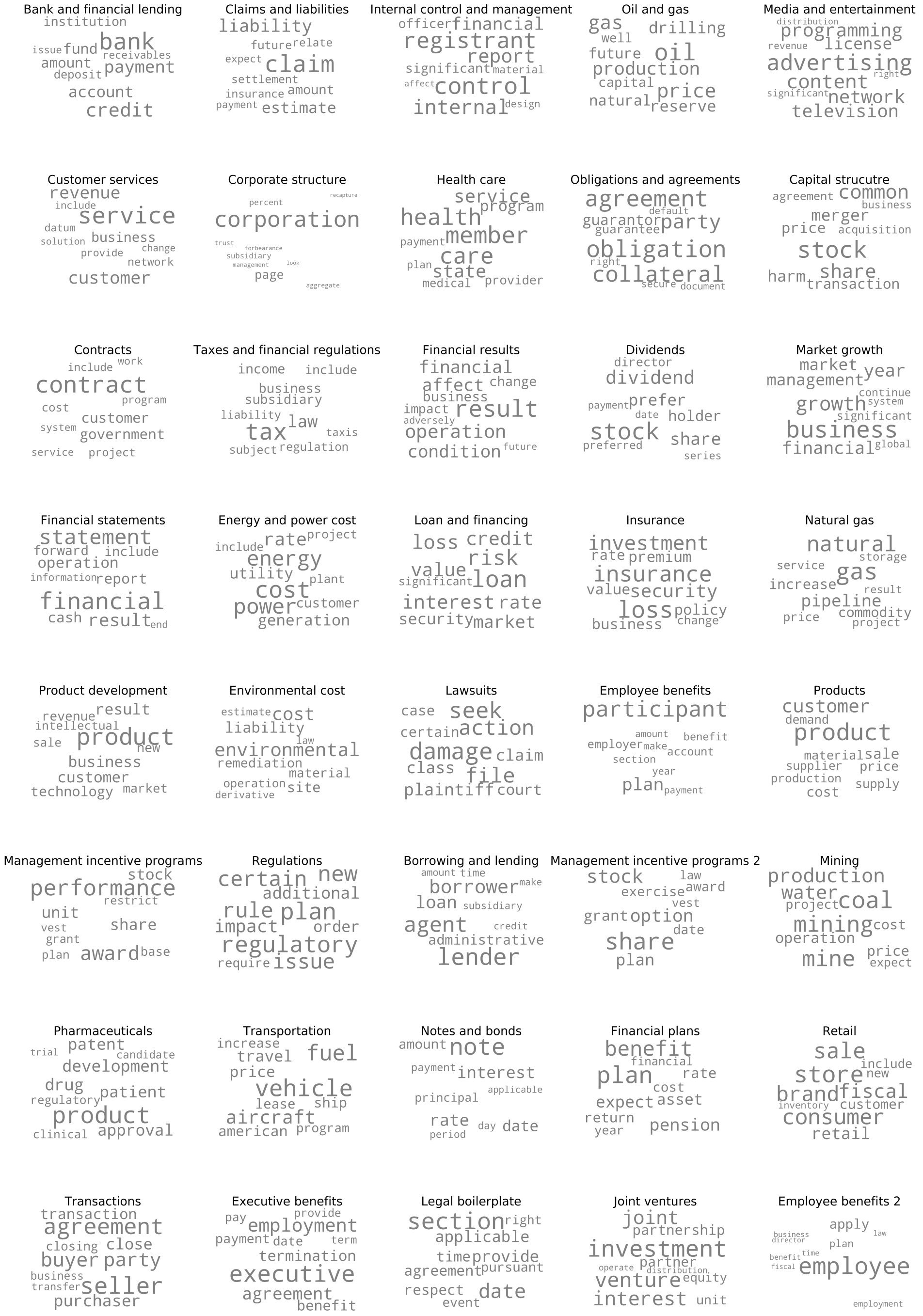


Figure 5.5: Word cloud for topics (1/2). Size of word indicate relative word importance in topic.

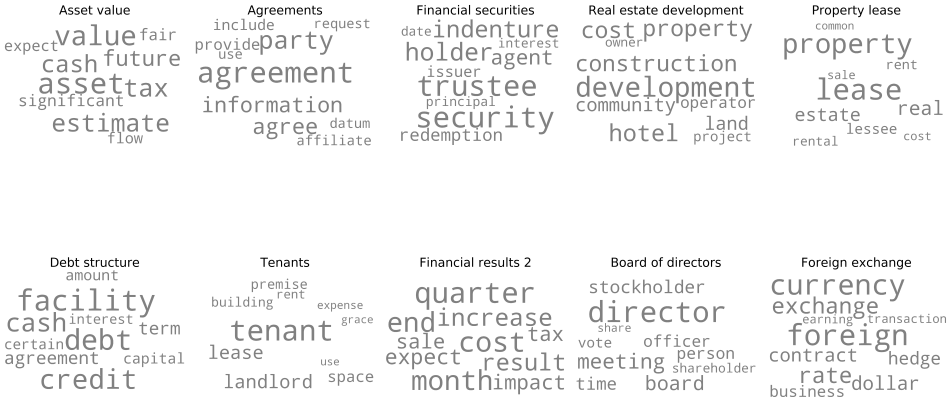


Figure 5.6: Word cloud for topics (2/2). Size of word indicate relative word importance in topic.

5.5 The Evolution of Topics in 10-Ks

Figure 5.7 illustrates the temporal change of topics in 10-Ks. We observe that the distribution varies with time, meaning that companies emphasize different topics from year to year. Dyer et al. (2017) investigate "*The evolution of 10-K textual disclosure*" using LDA to explain why such filings have increased in length the last years. With their topic distribution, they explain that three topics (*internal control, fair value, risk factor disclosures*) account for the increase, implying that companies tend to include long statements regarding these topics. Furthermore, the increased length can also be explained by regulatory requirements. Figure 5.7 is in line with several of the findings from citedyer$_{e}volution_{2}017.Forinstance, ourtopics$"*Internal control and management*"and"*Asset value*"$incr$

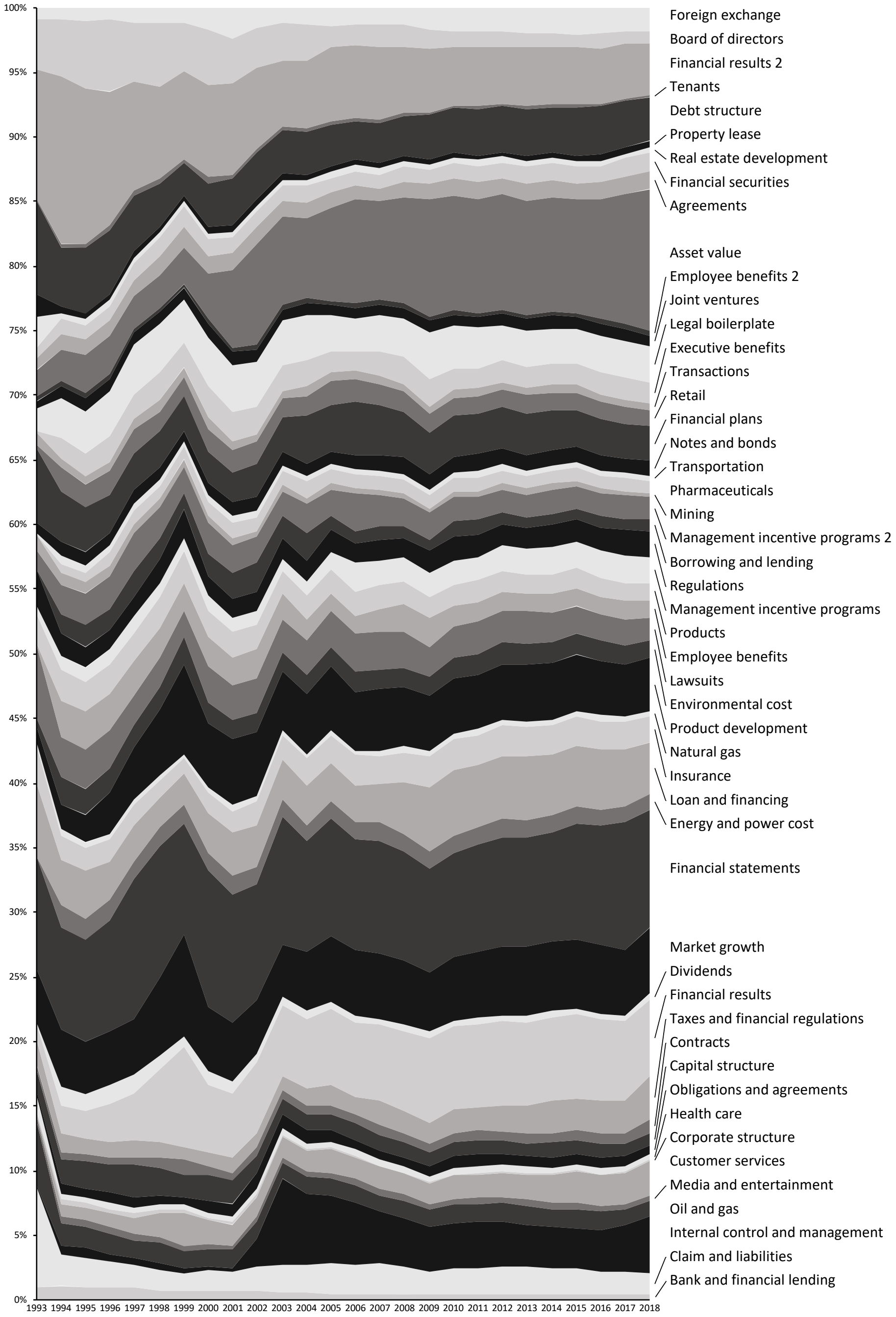


Figure 5.7: Average share of topics in 10-Ks per year

5.6 Method

There are three reasons why we cannot use the LDA output directly in the regressions. Firstly, the LDA algorithm outputs a topic distribution for each report that always will sum to one. Combined with control variables for years, and later also other variables, this quickly results in problems related to multicollinearity. Secondly, since we do not use control variables for sectors, using the distributions straight from the LDA algorithm causes the results to be mainly sector-specific (e.g., Oil and gas or Real estate). Lastly, as we show in Figure 5.7, the distributions of topics are highly dependent on which year the report was published.

Applying the same differencing scheme as we do for the sentiment- and readability variables takes care of the second and third problem, but does nothing to the multicollinearity. Consequently, we choose also to apply a non-linear transformation where we first normalize the variables (subtracting the mean and dividing by the standard deviation for each topic) and then map all values ≥ 2.5 to 1, values ≤ -2.5 to -1, and values between -2.5 and 2.5 to 0. Standard deviations of ± 2.5 correspond to roughly the 1% most extreme values.

The regression is then specified as follows:

$$return_{i,t,t+d} = \{\beta_n\}_{n=1}^{K}\{topic_{i,n,t}\}_{n=1}^{K} + \{\beta_y\}_{y=1994}^{2018}\{year_y\}_{y=1994}^{2018} + \epsilon_t \tag{5.8}$$

5.7 Results

Table 5.1 reports the results in terms of percentages for selected topics and holding periods d. The complete results for all topics is provided in Table A3.2 in the Appendix.

We see that the topic ”Health care” gives significant positive returns for up to 60 days after the reports are published. Health care may be related to both health care programs and the health care sector. However, due to the differencing scheme we have used, we deem a significant impact from the latter to be unlikely. Reports with increased discussions of environmental costs are associated with slightly positive abnormal returns, albeit the coefficient is only significant at 180 days. Increased discussions about ”Lawsuits” have the expected negative and significant coefficient in the short term, even though the effects are short-lived. Likewise, increased talk about ”Property lease” yields short-term negative abnormal returns. In the long term, the topic ”Financial plans” seem to be associated with positive and significant abnormal returns, despite exhibiting no short term effect.

When reading the inference from the coefficients in Table 5.1 (the full table consists of 350 coefficients), it is crucial to keep in mind that when using a $p < 0.01$ threshold for significance, one in one hundred coefficients will appear highly significant just by chance. We should, therefore, exercise caution and not overstate the results. However, as we show in Chapter 6, using the topics as a basis for a trading strategy yields statistically significant out-of-sample abnormal returns.

	Dependent variable: $return_t$						
	(1 days)	(5 days)	(10 days)	(30 days)	(60 days)	(180 days)	(252 days)
Health care	0.2844	**0.9981****	**1.3341*****	**1.9051*****	**2.4048****	3.5602*	**4.6153****
	(1.430)	(2.288)	(2.633)	(2.684)	(2.123)	(1.918)	(2.278)
Environmental cost	0.1791	0.4739*	**0.6638****	0.5088	1.7407*	**3.8489*****	2.8729*
	(1.296)	(1.959)	(2.153)	(0.741)	(1.901)	(2.735)	(1.750)
Lawsuits	**-0.4546****	**-0.6056****	-0.4026	-0.3175	0.1704	-0.182 8	-0.1063
	(-2.422)	(-2.106)	(-1.047)	(-0.528)	(0.170)	(-0.105)	(-0.052)
Financial plans	-0.0195	-0.1159	0.3549	1.1535*	**2.1646****	**4.0484****	**4.8328****
	(-0.098)	(-0.386)	(0.957)	(1.946)	(2.328)	(2.301)	(2.227)
Property lease	**-0.6485*****	**-0.7888****	-0.8708*	0.1011	-0.5063	0.4527	0.7673
	(-2.603)	(-2.466)	(-1.957)	(0.146)	(-0.487)	(0.252)	(0.353)
Foreign exchange	**-0.3267****	-0.1330	-0.5900*	**-1.5043****	-1.1981	**-3.0527****	-2.3909
	(-2.033)	(-0.521)	(-1.869)	(-2.018)	(-1.248)	(-2.164)	(-1.427)
Observations	9847	9750	9743	9721	9694	9568	9488
Note:					*p<0.1; **p<0.05; ***p<0.01		
					(t scores in parentheses)		

Table 5.1: Abnormal return regressed against the change in the portion of the annual report related to each topic. Table only shows selected topics, the rest is presented in Table A3.2. Regression is specified as $return_{i,t,t+d} = \{\beta_n\}_{n=1}^{K}\{topic_{n,t}\}_{n=1}^{K} + \{\beta_y\}_{y=1994}^{2018}\{year_y\}_{y=1994}^{2018} + \epsilon_t$ Variables are standardized, so coefficients can be interpreted as percentage abnormal return per standand deviation of abnormal topic discussion. Coefficients with p-values less than 0.05 are in bold.

Results for abnormal volume and volatility are included in Table A3.3 and Table A3.4 in the Appendix.

Chapter 6

Trading

In this chapter we consider the simple setting where an investor, due to regulations, skills, or cost limitations, needs to compose a portfolio consisting only of long positions. Can the patterns between sentiment, readability, and topics and subsequent returns be a useful input in this stock-picking problem?

For each year y, we select the reports from fiscal year $y-1$ that have been published by the end of March. This accounts for 69% of the reports. Based on the different sentiment-, readability- and topic variables from Chapter 4 and Chapter 5, we use a rolling prediction technique where we train a model from the previous five years ($y-6$ to $y-1$) and use the trained model to create an out-of-sample prediction for 252 trading days ahead in year y. We enter all positions on April 1 and exit all positions 252 trading days later. Then, we roll one year forward ($y \leftarrow y+1$) and continue the process. The stocks are Equally Weighted (EW), without rebalancing during the holding period. We begin in year $y = 2004$ as this is the first year where the sample size is large using differenced variables (differencing looses five years), and we end in our last year in the sample, $y = 2018$.

6.1 OLS as Prediction Model

We first implement the above strategy using Ordinary Least Squares (OLS) as the model on which to base the predictions. When using topics as prediction variables, we only use coefficients with corresponding p-values less than 10%. Since the analysis in Chapter 4 and Chapter 5 were done using linear OLS models, one could expect it to be possible to use the same model in a trading context.

Table 6.1 shows results for the portfolios based single variables, and a portfolio based on topics, regressed against the Fama-French factors. Although all variables on avarega generate positive annual returns, we observe that only $\Delta vader\_comp$, $\Delta vader\_neg$ and $topics$ have intercepts with some significance. The other portfolios attribute all their return to exposure to the known risk-factors, leaving the abnormal return close to 0.

	Trading strategy based upon:								
	$(\Delta naive\_tone)$	$(\Delta tfidf\_tone)$	$(\Delta vader\_pos)$	$(\Delta vader\_neg)$	$(\Delta vader\_comp)$	(Δnum_sents)	$(gunning\_fog)$	$(log\_filesize)$	$(topics)$
const	0.0121	0.0028	0.0019	0.0193**	0.0145*	0.0015	0.0056	0.0109	0.0158*
	(1.457)	(0.377)	(0.252)	(2.476)	(1.826)	(0.171)	(0.715)	(1.007)	(1.793)
Mkt-RF	0.9414***	0.9502***	1.0320***	1.0289***	1.0514***	1.0511***	0.9869***	0.9767***	1.0443***
	(91.120)	(93.981)	(92.394)	(99.411)	(102.566)	(74.817)	(83.540)	(71.495)	(84.183)
SMB	0.1781***	0.2080***	0.1658***	0.2319***	0.2275***	0.1799***	0.3295***	0.3713***	0.2504***
	(7.597)	(10.206)	(7.834)	(11.136)	(11.421)	(6.065)	(15.266)	(13.830)	(10.485)
HML	0.0987***	0.0870***	0.2316***	0.0838***	0.2019***	0.1196***	0.1011***	0.1235***	0.1734***
	(4.902)	(4.297)	(11.261)	(3.838)	(8.367)	(4.887)	(4.818)	(4.786)	(5.820)
R-squared	0.84	0.868	0.885	0.876	0.88	0.857	0.87	0.78	0.855
Intercept annual return	3.08%	0.72%	0.49%	4.98%**	3.72%*	0.37%	1.43%	2.78%	4.05%*
Observations	3646	3632	3646	3646	3646	3646	3646	3646	3646

Note: *p<0.1; **p<0.05; ***p<0.01
(t scores in parentheses)

Table 6.1: Fama-French regressions used as an evaluation of trading strategies. Daily return on Long portfolios is regressed on daily return on Fama-French factors. Regression is specified as $R_{p,t} - r_{f,t} = const + \beta_{MKT,i,t} \cdot (r_{MKT,t} - r_{f,t}) + \beta_{HML,i,t} \cdot r_{HML,t} + \beta_{SMB,i,t} \cdot r_{SMB,t}$. A significant constant indicates risk-adjusted (abnormal) return.

6.2 AdaBoost with CART Trees as Prediction Model

In practice, it is likely that the relationships between sentiment, readability, and topic content is more complicated and non-linear than which an OLS model allows for. We therefore implement a more advanced model to enable trading based on more complicated criteria.

We now base our predictions on a meta-estimator known as AdaBoost-SAMME[1] developed by Hastie et al. (2009). Weak learners are learning algorithms that are only slightly better than random guessing. AdaBoost (short for Adaptive Boosting), introduced by Freund and Schapire (1997), is a meta-algorithm based on combining the weighted classification of multiple weak learners into one combined classification. Each learner uses training data (attributes and classes) to build a model to be able to classify new, unseen instances into the correct classes. Adaboost trains the weak learners one by one, and each new learner adaptively focuses on the instances that the current classifiers collectively predict incorrectly. In Algorithm 1 in the Appendix, we provide a mathematical and algorithmic treatment of AdaBoost-SAMME.

For the individual classifiers, we use the non-parametric supervised learning method called Decision Tree Classifiers (Salzberg, 1994), more precisely a variation of C4.5 (Quinlan, 2014) called CART[2]. The combination of AdaBoost-SAMME and Decision Tree Classifiers has two very important hyperparameters; how many weak learners to use and the maximum height of each decision tree. After a crude grid search on a validation set consisting of $y = \{2004, 2005\}$, we achieve reasonable results using 200 weak learners and a maximum tree height of 15.

We use different subsets of the measures generated in Chapter 4 and Chapter 5 as attributes, divide the abnormal returns in the training set into five quintiles and use these as target classes. Each year on April 1, we enter long positions in the stocks that the algorithm predicts will end up in the upper quantile with regards to abnormal return, and hold these positions (without rebalancing) for 252 trading days. To verify that the returns are not resulting from exposure to the risk-factors in Fama and French (1993), we regress the daily excess returns of the portfolios on the common risk factors, and we consider any statistically significant intercept as proof of mispricing relative to the risk factors from Fama and French (1993).

We run the analysis for a selection of prediction variables. Since the rules for building the Decision Trees have some elements of randomness in it, the results could differ sligtly each run.[3] We therefore report the average values of 100 iterations.

To verify that the results are not influenced by selection bias, we include ”Comparable Market” portfolio. We generate this portfolio in the same way as the others, but where the model outputs buy signals on all available assets, i.e., creating an EW portfolio consisting of all stocks available in the dataset.

[1] SAMME is an abbreviation of ”Stagewise Additive Modeling using a Multi-class Exponential loss function”. This version of AdaBoost enables multi-class prediction, as opposed to the original AdaBoost algorithm, which is only able to predict two-class problems.

[2] CART is the Decision Tree version implemented by default in Python’s *scikit* library. It is very similar to C4.5, but supports numerical target variables (regression) and does not convert the tree into decision rules.

[3] The randomness has to do with the choices of which variables to split on in which order. Since the problem of making an optimal decision tree is NP complete, heuristics known as improvement criterions are used for this task. If two variables yield the same information criterion, a variable is selected at random.

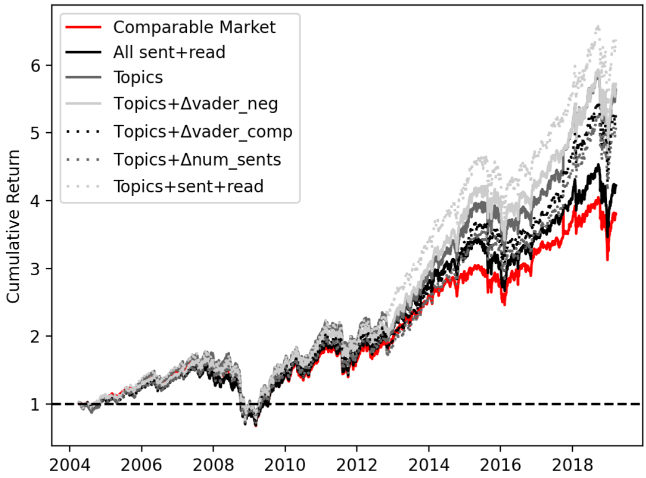


Figure 6.1: Cumulative return of AdaBoost portfolios (not risk-adjusted) compared to cumulative return of Comparable Market (not risk-adjusted)

Figure 6.1 contains a plot of raw cumulative returns (before any adjustment for exposure to Fama and French (1993) risk factors) together with the return of the ”Comparable Market” portfolio. The best AdaBoost portfolio outperforms the comparable market by about a factor of 1.5, and all AdaBoost portfolios have an overall return that is higher than the Comparable Market portfolio.

	Return (total)	Return (annualized)	Volatility (annualized)	Sharp Ratio (annual)
Comparable Market	277%	9.3%	21.1%	0.38
Sentiment + Readability	310%	9.9%	20.9%	0.41
Topics	489%	12.6%	21.9%	0.52
Topics + $\Delta vader\_neg$	472%	12.3%	21.3%	0.52
Topics + $\Delta vader\_comp$	437%	11.9%	21.3%	0.50
Topics + Δnum_sents	428%	11.7%	21.3%	0.49
Topics + Sentiment + Readability	494%	12.6%	21.2%	0.53

Table 6.2: Summary of AdaBoost portfolios. Values are calculated for 100 trails of each portfolio, before being averaged.

Table 6.2 shows a summary of each of the portfolios generated by the AdaBoost algorithm, before any adjustments for risk exposure. All portfolios clearly outperform the market portfolio in terms of return and Sharpe ratio, and the annual volatility is of comparable magnitude.

In Table 6.3, we show the regression results from daily portfolio returns regressed against the Fama and French (1993) risk factors. The portfolio based on $topics$ and $\Delta vader\_neg$ achieves an average abnormal daily return of 0.015%, equivalent to an annual abnormal return of 3.8%. This coefficient is statistically significant at a 1% level. Furthermore, $topics$ alone, $topics$ combined with $\Delta vader\_comp$, and $topics$ combined with Δnum_sents, achieve risk-adjusted annual returns of between 3.1% and 3.6% annually. We note that the portfolios based merely on sentiment- and readability measures are outperformed by all portfolios based on topics. However, including more variables is not always better, as the combination of all available variables is still inferior to using only topics and $\Delta vader\_neg$. This is contrary to what one might infer from Table 6.3, where the best portfolio in terms of Sharpe ratio (and return) is the portfolio generated with all available variables (althoug the differences are small). From the intercept for "Comparable Market" in Table 6.3, we see that it is not significantly different from zero, which is what we would expect.

Figure 6.2 is a plot of the cumulative daily abnormal return, which we define as the sum of the intercepts and the residuals from the Fama and French (1993) regression. For comparison, we also include the cumulative Fama and French (1993) risk factors. From this plot, we can see that the algorithm seems to work well until 2016, and that the risk-adjusted return is slightly negative between 2016 and 2018.

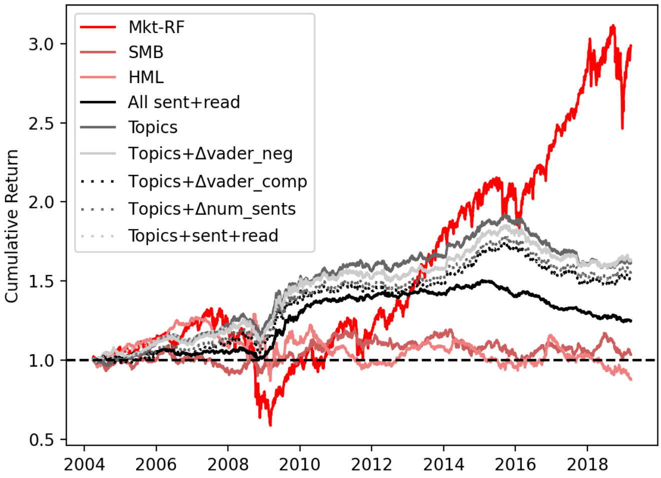


Figure 6.2: Cumulative abnormal return (intercept+residuals from Fama-French regressions) for AdaBoost portfolios compared to cumulative return of Fama-French factors

For completeness, we also include portfolios generated from one variable at a time. As we can see in Table 6.4, the intercepts are all insignificantly different from zero, indicating that for these particular settings, we are better of using standard OLS. This should be of no surprise, since one-dimentional problem spaces do not do justice to a boosting algorithm like AdaBoost. In particular when the data is noisy, and the hyperparameters are optimized for the multiple predictor setting (200 classifiers, each with maximum height of 15), the

	Dependent variable: Portfolio Return						
	(Comparable Market)	(All sent.+read.)	(Topics)	(Topics+$\Delta vader\_neg$)	(Topics+$\Delta vader\_comp$)	(Topics+Δnum_sents)	(Topics+sent.+read.)
const	0.0001	0.0042	0.0140***	0.0149***	0.0123**	0.0137***	0.0146***
	(0.024)	(0.933)	(2.783)	(2.973)	(2.514)	(2.775)	(3.024)
Mkt-RF	1.0478***	1.0333***	1.1127***	1.0871***	1.0882***	1.0869***	1.0849***
	(207.224)	(173.336)	(170.679)	(169.708)	(168.964)	(170.154)	(168.064)
SMB	0.1953***	0.2443***	0.2798***	0.2667***	0.2655***	0.2677***	0.2649***
	(15.688)	(20.608)	(20.553)	(19.011)	(18.929)	(19.202)	(18.661)
HML	0.2154***	0.1157***	-0.0351**	-0.0222	-0.0220	-0.0227	-0.0228*
	(16.902)	(9.596)	(-2.534)	(-1.594)	(-1.571)	(-1.643)	(-1.654)
R-squared	0.958	0.955	0.95	0.949	0.951	0.95	0.952
Intercept annual return	0.03%	1.07%	3.60%	3.83%	3.15%	3.51%	3.74%
Observations	3646	3646	3646	3646	3646	3646	3646
Note:							*p<0.1; **p<0.05; ***p<0.01 *(t scores in parentheses)*

Table 6.3: Daily return on AdaBoost portfolios regressed on daily return on Fama-French factors. Regression is specified as $R_{p,t} - r_{f,t} = const + \beta_{MKT,i,t} \cdot (r_{MKT,t} - r_{f,t}) + \beta_{HML,i,t} \cdot r_{HML,t} + \beta_{SMB,i,t} \cdot r_{SMB,t}$ A significant constant indicates risk-adjusted (abnormal) return.

poor performance is unsurprising. In general, good classifiers should capture patterns in the training data, and at the same time generalize adequately to be able to infer new instances. With single variables as predictors, noisy data and suboptimal hyperparameters, it is likely that the classifiers overfit the training data, hence being ineffective in classifying the test data (in this case doing the actual trading).

	Dependent variable: Portfolio Alpha						
	$(\Delta naive\_tone)$	$(\Delta vader\_pos)$	$(\Delta vader\_neg)$	$(\Delta vader\_comp)$	(Δnum_sents)	$(gunning\_fog)$	$(log\_filesize)$
const	-0.0021	0.0071	-0.0026	0.0011	0.0089	0.0046	0.0068
	(-0.436)	(1.080)	(-0.439)	(0.214)	(1.373)	(0.720)	(1.406)
Mkt-RF	1.0272***	1.0397***	1.0444***	1.0377***	1.0431***	1.0597***	1.0398***
	(150.726)	(143.703)	(179.728)	(146.456)	(152.205)	(163.705)	(154.773)
SMB	0.2342***	0.2157***	0.1899***	0.2208***	0.2216***	0.2422***	0.2768***
	(17.133)	(14.055)	(12.314)	(14.580)	(13.917)	(14.831)	(19.756)
HML	0.1607***	0.1583***	0.1259***	0.1293***	0.2471***	0.1584***	0.1525***
	(11.451)	(11.751)	(9.909)	(8.099)	(16.876)	(12.812)	(9.433)
R-squared	0.951	0.912	0.925	0.939	0.917	0.92	0.951
Intercept annual return	-0.52%	1.81%	-0.66%	0.29%	2.27%	1.16%	1.73%
Observations	3646	3646	3646	3646	3646	3646	3646
Note:							*p<0.1; **p<0.05; ***p<0.01 *(t scores in parentheses)*

Table 6.4: Daily return on AdaBoost portfolios regressed on daily return on Fama-French factors. Regression is specified as $R_{p,t} - r_{f,t} = const + \beta_{MKT,i,t} \cdot (r_{MKT,t} - r_{f,t}) + \beta_{HML,i,t} \cdot r_{HML,t} + \beta_{SMB,i,t} \cdot r_{SMB,t}$ A significant constant indicates risk-adjusted (abnormal) return.

We want to emphasize that the trading is done only on the long term effects of the company reports. Since we do not enter any position before April 1, weeks and months may have passed since the report was published, and we decide to buy the stock. Contrary to many other algorithmic trading procedures, this procedure does not require any form of low-latency access to financial markets. Neither is it particularly vulnerable to trading costs since the strategy requires only buying and selling stocks from between 50 and 100 companies each year.

Financial data is known to be noisy, and there are of course more important factors (omitted variables) that determine long term abnormal return other than the content of the annual report. The AdaBoost-SAMME algorithm yields the above results even though the exponential loss function used by AdaBoost often tends to make the algorithm vulnerable to noisy data. Although we tested a small selection of other meta algorithms and weak learners, we have by no means exhausted the vast amount of available options. Other algorithms may be able to improve the results. One particularly interesting meta algorithm in this respect is Brown-boosting, which is found to be better at handling noisy data than AdaBoost, while retaining all of the other important properties (Mcdonald et al., 2004).

We need to address one slight caveat regarding the use of topics in our trading model. Although only the report itself, together with the LDA model, is utilized when calculating a report's topic distribution, we use reports from all years to generate the LDA model. Hence, information that should, strictly speaking, not be available at the time of trading, is used. The likelihood that this affects the results is practically minimal, however. The LDA model generation is not guided by any financial information, such as abnormal returns. Furthermore, the LDA model changes little from one time period to another. Generating a new LDA model on a rolling basis would solve this theoretical issue, but would each generate slightly different topics from the ones we analyze in Chapter 5. Additionally, each LDA model takes about three days to build, so time cost makes this option unattractive.

In theory, we should be able to create hedge portfolios (zero-cost portfolios), where we use the proceeds from the short positions to buy long positions. Assuming the samples in the hedge portfolios are representative in their exposure to risk factors, the net exposure would be close to zero, and any return (or loss) could be interpreted as a risk-adjusted abnormal return (loss). We are, however, unable to achieve significant returns using this method, resulting from the model's lack of ability to select good short candidates. Indeed, another model may have been more appropriate for this purpose.

Chapter 7

Conclusion

Using a large sample of $15,700$ annual reports from S&P 500 companies in the period from 1994 to 2018, we identify systematic relationships between the reports' sentiment, readability and topics addressed and subsequent long-term abnormal returns. The main contributions from this thesis are twofold; we develop an enhanced measure for the sentiment of annual reports, and we identify relationships between topics addressed in the reports and subsequent abnormal returns.

We use our modified VADER measure of negative sentiment to find that when the negativity of the sentiment in report is one standard deviation above average, subsequent abnormal one-year return is -1.34%. However, the degree of positivity of a report is not significantly related to abnormal return. We offer a "masking" explanation to this observation; negative messages are often padded with positivity to dampen the negative information's effect. Interestingly, we also observe that the delay between the publishing time and the materialization of abnormal returns is longspun, in most cases up to one year, indicating that financial markets are ineffective in absorbing the textual information contained in the reports. Our modified VADER measure appears to capture these effects better than the measures traditionally implemented to capture annual report sentiment.

We also find that a reports' readability is positively related to subsequent one-year abnormal returns, both measured in terms of the Fog index and of the natural logarithm of the reports' file sizes. We echo the explanations of Li (2008) and Bloomfield (2008), who attribute this observation to firms using longer reports to conceal damaging information strategically, and poorly performing firms needing to explain the company situation more thoroughly. The finding may indicate that executives wanting to maximize company valuations should produce annual reports that are clear and easily readable in order to minimize the risk associated with the companies' equity. However, although the observations are statistically significant, most of the identified effects only explain about 1% of the variation in abnormal returns. Therefore, the methods we use should only be applied as a supplement to the traditional ways of interpreting the annual reports. It might also be that the measures we use are proxies for other contemporaneous information and that this is the real cause behind the findings.

We find that short-term volatility and trading volume decreases significantly with report readability, but remains unaffected by report sentiment. The effects on volume and volatility do, in most cases, last for about one week. We suppose a plausible explanation is that unclear reports make investors disagree about how to interpret the information, and that this, in turn, leads to higher trading volume and volatility.

Through the use of LDA, we identify 50 topics commonly appearing in forward-looking statements in annual reports. We then analyze how the prevalence of these topis is related to subsequent abnormal returns. Our results suggest that increased forward-looking discussions about the topics ”Health care”, ”Environmental cost”, and ”Financial plans” are related to subsequent positive abnormal returns, while ”Lawsuits”, ”Property lease”, and ”Foreign exchange” tend to precede negative abnormal returns.

Finally, we implement a trading strategy based on AdaBoost and CART trees that takes as inputs the sentiment-, readability- and topic measures from the reports, and outputs the uppermost quintile in terms of next year's predicted abnormal return. After controlling for exposure to the Fama and French (1993) risk factors, we are able to make profitable investment strategies that achieve an annualized abnormal return of up to 3.8% per year.

Although our trading strategy performs adequately, further investigation can be conducted to enhance the trading model by using different learning algorithms and parameters, different combinations of variables, or different ways to define the variables. It would also be of great interest to inspect the generated models' inner workings to identify which patterns it exploits to classify the stocks. One extension already mentioned is to replace AdaBoost with Brown-boosting. Furthermore, we expect more advanced methods for natural language processing, such as Recurrent Neural Networks (RNN), to be even better at identifying relationships between textual semantics in annual reports and subsequent financial performance, but with the downside that the results of these models are even harder to explain.

Bibliography

Aletras, N., Stevenson, M., 2013. Evaluating Topic Coherence Using Distributional Semantics, in: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, Association for Computational Linguistics, Potsdam, Germany. pp. 13–22. URL: https://www.aclweb.org/anthology/W13-0102.

Asthana, S., Balsam, S., 2001. The effect of EDGAR on the market reaction to 10-K filings. Journal of Accounting and Public Policy 20, 349–372. URL: http://www.sciencedirect.com/science/article/pii/S0278425401000357, doi:10.1016/S0278-4254(01)00035-7.

Ball, C., Hoberg, G., Maksimovic, V., 2013. Disclosure, Business Change and Earnings Quality. SSRN Scholarly Paper ID 2260371. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=2260371, doi:10.2139/ssrn.2260371.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022. URL: http://www.jmlr.org/papers/v3/blei03a.

Bloomfield, R., 2008. Discussion of "Annual report readability, current earnings, and earnings persistence". Journal of Accounting and Economics 45, 248–252. URL: http://www.sciencedirect.com/science/article/pii/S0165410108000190, doi:10.1016/j.jacceco.2008.04.002.

Bouma, G., 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction , 11.

Brown, N.C., Crowley, R.M., Elliott, W.B., 2017. What Are You Saying? Using topic to Detect Financial Misreporting. Journal of Accounting Research 58, 237–291. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12294, doi:10.1111/1475-679X.12294.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.M., 2009. Reading Tea Leaves: How Humans Interpret Topic Models , 9.

Chen, H., De, P., Hu, Y.J., Hwang, B.H., 2013. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. SSRN Scholarly Paper ID 1807265. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=1807265.

Cohen, L., Malloy, C., Nguyen, Q., 2018. Lazy Prices. Working Paper 25084. National Bureau of Economic Research. URL: http://www.nber.org/papers/w25084, doi:10.3386/w25084.

Davis, A.K., Ge, W., Matsumoto, D.A., Zhang, J.L., 2014. The Effect of Manager-Specific Optimism on the Tone of Earnings Conference Calls. SSRN Scholarly Paper ID 1982259. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=1982259.

Davis, A.K., Piger, J., Sedor, L.M., 2011. Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language. SSRN Scholarly Paper ID 875399. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=875399.

Davis, A.K., Tama-Sweet, I., 2011. Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases Versus MD&A. SSRN Scholarly Paper ID 1866369. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=1866369.

Doran, J.S., Peterson, D.R., Price, S.M., 2012. Earnings Conference Call Content and Stock Price: The Case of REITs. The Journal of Real Estate Finance and Economics 45, 402–434. URL: https://doi.org/10.1007/s11146-010-9266-z, doi:10.1007/s11146-010-9266-z.

Dyer, T., Lang, M., Stice-Lawrence, L., 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. Journal of Accounting and Economics 64, 221–245. URL: http://www.sciencedirect.com/science/article/pii/S0165410117300484, doi:10.1016/j.jacceco.2017.07.002.

Easley, D., O'Hara, M., 2004. Information and the Cost of Capital. The Journal of Finance 59, 1553–1583. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2004.00672.x, doi:10.1111/j.1540-6261.2004.00672.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2004.00672.x.

Easton, P.D., Zmijewski, M.E., 1993. SEC Form 10K/10Q Reports and Annual Reports to Shareholders: Reporting Lags and Squared Market Model Prediction Errors. Journal of Accounting Research 31, 113–129. URL: https://www.jstor.org/stable/2491044, doi:10.2307/2491044. publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley].

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56. URL: http://www.sciencedirect.com/science/article/pii/0304405X93900235, doi:10.1016/0304-405X(93)90023-5.

Feuerriegel, S., Pröllochs, N., 2018. Investor Reaction to Financial Disclosures across Topics: An Application of Latent Dirichlet Allocation. Decision Sciences n/a. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/deci.12346, doi:10.1111/deci.12346. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/deci.12346.

Feuerriegel, S., Ratku, A., Neumann, D., 2016. Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation, in: 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 1072–1081. doi:10.1109/HICSS.2016.137. iSSN: 1530-1605.

Franco, G.D., Hope, O.K., Vyas, D., Zhou, Y., 2015. Analyst Report Readability. Contemporary Accounting Research 32, 76–104. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12062, doi:10.1111/1911-3846.12062. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1911-3846.12062.

Freund, Y., Schapire, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55, 119–139. URL: http://www.sciencedirect.com/science/article/pii/S002200009791504X, doi:10.1006/jcss.1997.1504.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101, 5228–5235. URL: https://www.pnas.org/content/101/suppl_1/5228, doi:10.1073/pnas.0307752101. publisher: National Academy of Sciences Section: Colloquium.

Guay, W., Samuels, D., Taylor, D., 2016. Guiding through the Fog: Financial statement complexity and voluntary disclosure. Journal of Accounting and Economics 62, 234–269. URL: http://www.sciencedirect.com/science/article/pii/S0165410116300489, doi:10.1016/j.jacceco.2016.09.001.

Guo, L., Shi, F., Tu, J., 2016. Textual analysis and machine leaning: Crack unstructured data in finance and accounting. The Journal of Finance and Data Science 2, 153–170. URL: http://www.sciencedirect.com/science/article/pii/S2405918816300496, doi:10.1016/j.jfds.2017.02.001.

Hastie, T., Rosset, S., Zhu, J., Zou, H., 2009. Multi-class AdaBoost. Statistics and Its Interface 2, 349–360. URL: http://www.intlpress.com/site/pub/pages/journals/items/sii/content/vols/0002/0003/a008/, doi:10.4310/SII.2009.v2.n3.a8.

Henry, E., 2008. Are Investors Influenced By How Earnings Press Releases Are Written? The Journal of Business Communication (1973) 45, 363–407. URL: https://journals.sagepub.com/doi/abs/10.1177/0021943608319388, doi:10.1177/0021943608319388.

Hilary, G., Biddle, G.C., Verdi, R.S., 2009. How Does Financial Reporting Quality Relate to Investment Efficiency? Technical Report hal-00481731. HAL. URL: https://ideas.repec.org/p/hal/journl/hal-00481731.html.

Hoberg, G., Lewis, C., 2017. Do fraudulent firms produce abnormal disclosure? Journal of Corporate Finance 43, 58–85. URL: http://www.sciencedirect.com/science/article/pii/S0929119916303637, doi:10.1016/j.jcorpfin.2016.12.007.

Hutto, C.J., Gilbert, E., 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text, in: Eighth International AAAI Conference on Weblogs and Social Media. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

Hüfner, B., 2007. The SEC's MD&A: Does it Meet the Informational Demands of Investors? Schmalenbach Business Review 59, 58–84. URL: https://doi.org/10.1007/BF03396742, doi:10.1007/BF03396742.

Jiang, F., Lee, J., Martin, X., Zhou, G., 2019. Manager sentiment and stock returns. Journal of Financial Economics 132, 126–149. URL: http://www.sciencedirect.com/science/article/pii/S0304405X18302770, doi:10.1016/j.jfineco.2018.10.001.

Karapandza, R., 2016. Stock returns and future tense language in 10-K reports. Journal of Banking & Finance 71, 50–61. URL: http://www.sciencedirect.com/science/article/pii/S0378426616300577, doi:10.1016/j.jbankfin.2016.04.025.

Kearney, C., Liu, S., 2014. Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis 33, 171–185. URL: http://www.sciencedirect.com/science/article/pii/S1057521914000295, doi:10.1016/j.irfa.2014.02.006.

Kothari, S.P., Li, X., Short, J.E., 2009. The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. The Accounting Review 84, 1639–1670. URL: https://aaajournals.org/doi/10.2308/accr.2009.84.5.1639, doi:10.2308/accr.2009.84.5.1639.

Kumar, B.S., Ravi, V., 2016. A survey of the applications of text mining in financial domain. Knowledge-Based Systems 114, 128–147. URL: http://www.sciencedirect.com/science/article/pii/S0950705116303872, doi:10.1016/j.knosys.2016.10.003.

Lawrence, A., 2013. Individual investors and financial disclosure. Journal of Accounting and Economics 56, 130–147. URL: http://www.

sciencedirect.com/science/article/pii/S0165410113000359, doi:10.1016/j.jacceco.2013.05.001.

Lewis, N.R., Parker, L.D., Pound, G.D., Sutcliffe, P., 1986. Accounting Report Readability: The Use of Readability Techniques. Accounting and Business Research 16, 199–213. URL: https://doi.org/10.1080/00014788.1986.9729318, doi:10.1080/00014788.1986.9729318.

Li, F., 2006. Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? SSRN Scholarly Paper ID 898181. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=898181.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221–247. URL: http://www.sciencedirect.com/science/article/pii/S0165410108000141, doi:10.1016/j.jacceco.2008.02.003.

Li, F., 2010. The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. Journal of Accounting Research 48, 1049–1102. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-679X.2010.00382.x, doi:10.1111/j.1475-679X.2010.00382.x.

Loughran, T., McDonald, B., 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66, 35–65. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x, doi:10.1111/j.1540-6261.2010.01625.x.

Loughran, T., McDonald, B., 2014. Measuring Readability in Financial Disclosures. The Journal of Finance 69, 1643–1671. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12162, doi:10.1111/jofi.12162.

Loughran, T., McDonald, B., 2015. The Use of Word Lists in Textual Analysis. Journal of Behavioral Finance 16, 1–11. URL: http://www.tandfonline.com/doi/abs/10.1080/15427560.2015.1000335, doi:10.1080/15427560.2015.1000335.

Loughran, T., McDonald, B., 2016. Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research 54, 1187–1230. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12123, doi:10.1111/1475-679X.12123.

Loughran, T., McDonald, B., 2019. Textual Analysis in Finance. SSRN Scholarly Paper ID 3470272. Social Science Research Network. Rochester, NY. URL: https://papers.ssrn.com/abstract=3470272.

McDonald, B., Loughran, T., 2015. Textual Analysis in Accounting and Finance: A Survey - LOUGHRAN - 2016 - Journal of Accounting Research - Wiley Online Library. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/1475-679X.12123.

Mcdonald, R.A., Hand, D.J., Eckley, I.A., 2004. A MULTICLASS EXTENSION TO THE BROWNBOOST ALGORITHM. International Journal of Pattern Recognition and Artificial Intelligence 18, 905–931. URL: https://www.worldscientific.com/doi/abs/10.1142/S0218001404003472, doi:10.1142/S0218001404003472.

Miller, B.P., 2010. The Effects of Reporting Complexity on Small and Large Investor Trading. The Accounting Review 85, 2107–2143. URL: https://www.aaajournals.org/doi/abs/10.2308/accr.00000001, doi:10.2308/accr.00000001.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, United Kingdom. pp. 262–272.

Molnár, P., 2012. Properties of range-based volatility estimators. International Review of Financial Analysis 23, 20–29. URL: http://www.sciencedirect.com/science/article/pii/S1057521911000731, doi:10.1016/j.irfa.2011.06.012.

Newman, D., Lau, J.H., Grieser, K., Baldwin, T., 2010. Automatic evaluation of topic coherence, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, California. pp. 100–108.

Price, S.M., Doran, J.S., Peterson, D.R., Bliss, B.A., 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance 36, 992–1011. URL: http://www.sciencedirect.com/science/article/pii/S0378426611002901, doi:10.1016/j.jbankfin.2011.10.013.

Quinlan, J.R., 2014. C4.5: Programs for Machine Learning. Elsevier. Google-Books-ID: b3ujBQAAQBAJ.

Rogers, J.L., Van Buskirk, A., Zechman, S.L.C., 2011. Disclosure Tone and Shareholder Litigation. The Accounting Review 86, 2155–2183. URL: https://www.aaajournals.org/doi/abs/10.2308/accr-10137, doi:10.2308/accr-10137.

Röder, M., Both, A., Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures | Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. URL: https://dl.acm.org/doi/abs/10.1145/2684822.2685324. archive Location: world Library Catalog: dl.acm.org.

Salzberg, S.L., 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning 16, 235–240. URL: http://link.springer.com/10.1007/BF00993309, doi:10.1007/BF00993309.

SEC, 2003. Interpretation: Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations; Release Nos. 33-8350; 34-48960; FR-72; December 19, 2003. URL: https://www.sec.gov/rules/interp/33-8350.htm.

Sohangir, S., Petty, N., Wang, D., 2018. Financial Sentiment Lexicon Analysis, in: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), pp. 286–289. doi:10.1109/ICSC.2018.00052.

Tennyson, B.M., Ingram, R.W., Dugan, M.T., 1990. Assessing the Information Content of Narrative Disclosures in Explaining Bankruptcy. Journal of Business Finance & Accounting 17, 391–410. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-5957.1990.tb01193.x, doi:10.1111/j.1468-5957.1990.tb01193.x.

Tetlock, P.C., 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance 62, 1139–1168. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2007.01232.x, doi:10.1111/j.1540-6261.2007.01232.x.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance 63, 1437–1467. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2008.01362.x, doi:10.1111/j.1540-6261.2008.01362.x.

Tripathi, P., Vishwakarma, S.K., Lala, A., 2015. Sentiment Analysis of English Tweets Using Rapid Miner, in: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 668–672. doi:10.1109/CICN.2015.137. iSSN: 2472-7555.

Twedt, B., Rees, L., 2012. Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. Journal of Accounting and Public Policy 31, 1–21. URL: http://www.sciencedirect.com/science/article/pii/S0278425411001189, doi:10.1016/j.jaccpubpol.2011.10.010.

Wang, C.J., Tsai, M.F., Liu, T., Chang, C.T., 2013. Financial Sentiment Analysis for Risk Prediction, in: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Nagoya, Japan. pp. 802–808. URL: https://www.aclweb.org/anthology/I13-1097.

You, H., Zhang, X.j., 2009. Financial reporting complexity and investor underreaction to 10-K information. Review of Accounting Studies 14, 559–586. URL: https://doi.org/10.1007/s11142-008-9083-2, doi:10.1007/s11142-008-9083-2.

Appendix

Appendix A1

Literature Study

A1.1 Mentioned Lexicons

A1.1.1 The Affective Norms for English Words Lexicon (ANEW)

ANEW consists of 1,034 English words, which have been ranked by intensity. The lexicons we discussed in subsection 2.2.1 do not consider intensity of sentiment language, but rather classifies word sentiment as a binary; either positive or negative. The ANEW lexicon, on the other hand, classify words with a sentiment valence between 1-9, where 1 is most negative, 9 is most positive and 5 is considered to be neutral. For example, the word *bankrupt* has a Valence of 2.00, *mistake* has a valance of 2.86, *improve* a valence of 7.65 and paradise a valance of 8.72.

A1.1.2 Linguistic Inquiry and Word Count (LIWC)

LIWC was developed in the early 1990s, and uses 4,500 words and word stems to calculate the percentage of words that match particular feelings, emotions and thinking styles from a text. The LIWC dictionary contains 55 word categories, and one single word may be a part of several categories. For instance, the word ”happy” and ”cried” belong to the categories ”positive emotions” and ”negative emotions”, respectively.

Appendix A2

The Impact of Sentiment and Readability

A2.1 Results 2004-2018

Table A2.1, Table A2.2, Table A2.3 are the same regressions as presented in Chapter 4, but using reports from the time period 2004-2018 instead of 1994-2018.

	Dependent variable: return$_t$								
	(1 days)	(3 days)	(5 days)	(10 days)	(30 days)	(45 days)	(60 days)	(180 days)	(252 days)
$\Delta naive\_tone$	0.0567	0.0308	0.0424	-0.0000	0.0919	0.0403	0.1085	0.6118^{*}	0.6801^{*}
	(1.620)	(0.669)	(0.815)	(-0.000)	(0.923)	(0.302)	(0.696)	(1.877)	(1.696)
$\Delta tfidf\_tone$	0.0169	0.0322	0.0557	0.0084	0.0226	-0.1014	0.0488	0.4307	0.4850
	(0.435)	(0.660)	(1.002)	(0.124)	(0.229)	(-0.752)	(0.308)	(1.341)	(1.120)
$\Delta vader\_pos$	0.0328	0.0277	0.0104	0.0244	0.0651	0.1577	0.2786^{*}	0.3764	0.6014
	(0.866)	(0.552)	(0.185)	(0.360)	(0.631)	(1.184)	(1.800)	(1.153)	(1.310)
$\Delta vader\_neg$	-0.0279	-0.0370	0.0212	-0.0283	-0.0475	0.0179	-0.2486	-0.8418^{**}	-1.2186^{***}
	(-0.590)	(-0.599)	(0.336)	(-0.376)	(-0.434)	(0.120)	(-1.425)	(-2.346)	(-2.824)
$\Delta vader\_comp$	0.0531	0.0422	-0.0050	0.0386	0.0855	0.0339	0.2830^{*}	0.7933^{**}	1.1416^{***}
	(1.330)	(0.826)	(-0.085)	(0.547)	(0.800)	(0.233)	(1.691)	(2.372)	(2.721)
Δnum_sents	-0.0064	-0.0299	-0.0222	-0.0294	-0.0784	-0.0665	-0.0201	-0.5946^{*}	-0.6626^{*}
	(-0.121)	(-0.462)	(-0.298)	(-0.325)	(-0.543)	(-0.388)	(-0.110)	(-1.856)	(-1.768)
$gunning\_fog$	-0.0406	-0.0675	0.0066	0.0184	-0.0694	0.1114	0.0959	-0.4348	-0.2617
	(-1.107)	(-1.334)	(0.123)	(0.278)	(-0.684)	(0.869)	(0.643)	(-1.429)	(-0.564)
$log\_filesize$	-0.0950^{**}	-0.1216^{**}	-0.0788	-0.1160^{*}	-0.1626	-0.0946	-0.0758	-0.7751^{**}	-1.1122^{**}
	(-2.298)	(-2.100)	(-1.385)	(-1.668)	(-1.474)	(-0.698)	(-0.495)	(-2.346)	(-2.082)
Observations	9937	9619	9617	9610	9590	9577	9567	9440	9362
Δ Observations	8784	8697	8695	8688	8668	8656	8646	8530	8455

Note: *(t scores in parentheses)* *p<0.1; **p<0.05; ***p<0.01

Table A2.1: Impact of sentiment and readability on abnormal return (2004-2018). Regression is specified as $return_{i,t,d} = \beta\overline{X_{i,t}} + \{year_y\}_{y=2004}^{2018} + \epsilon$, where $X_{i,t}$ takes any of the independent variables. The regressions are performed on each dependent variable individually. Table shows the beta coefficients with its respective t-scores. Each row and column are separate regressions, where the rows represent the independent variable and the columns represent the time horizon of the dependent variable. Δ Observations is number of observations for differenced variables (name starts with "Δ "), while Observations is number of observations for the other regressions.

	Dependent variable: volatility$_t$										
	(0 days)	(1 days)	(2 days)	(3 days)	(4 days)	(5 days)	(6 days)	(7 days)	(8 days)	(9 days)	(10 days)
$\Delta naive\_tone$	-0.4797	0.8966	1.0011	-0.5497	-0.0184	0.3958	0.2475	-0.7623	-0.3969	-0.2851	-1.7121*
	(-0.473)	(0.960)	(1.148)	(-0.641)	(-0.022)	(0.451)	(0.290)	(-0.882)	(-0.447)	(-0.312)	(-1.960)
$\Delta tfidf\_tone$	0.5020	0.3808	0.6345	0.4716	0.5078	0.3359	-0.2601	0.2081	0.2767	-0.1895	0.0707
	(1.023)	(0.750)	(1.316)	(1.016)	(1.117)	(0.725)	(-0.564)	(0.439)	(0.599)	(-0.387)	(0.142)
$\Delta vader\_pos$	-0.9736	0.4269	0.5411	0.5838	0.1019	0.2670	-0.6806	-0.3076	-0.0435	-1.2272	-2.3284***
	(-0.920)	(0.443)	(0.641)	(0.703)	(0.119)	(0.302)	(-0.792)	(-0.353)	(-0.050)	(-1.391)	(-2.669)
$\Delta vader\_neg$	-0.5988	-0.5386	-0.8561	-0.9458	-0.0534	-0.3942	-1.1897	-0.7398	-0.8408	-0.2484	0.1334
	(-0.541)	(-0.522)	(-0.948)	(-1.078)	(-0.059)	(-0.427)	(-1.350)	(-0.831)	(-0.923)	(-0.273)	(0.145)
$\Delta vader\_comp$	0.0487	1.0050	0.5482	0.5453	0.0086	-0.0576	0.4783	0.3583	0.7410	-0.3290	-1.3782
	(0.046)	(1.040)	(0.668)	(0.643)	(0.010)	(-0.065)	(0.551)	(0.419)	(0.833)	(-0.376)	(-1.558)
Δnum_sents	0.4756	1.8264*	1.1189	1.7340**	2.0241**	0.8240	0.3085	1.8326**	1.6958*	0.3181	2.0245**
	(0.434)	(1.824)	(1.159)	(2.051)	(2.400)	(1.009)	(0.381)	(2.213)	(1.861)	(0.391)	(2.280)
$gunning\_fog$	1.9023**	2.0818**	1.9499**	3.0471***	2.7823***	1.5570*	0.4431	2.0403**	1.6151**	1.5557*	2.4980***
	(1.972)	(2.370)	(2.473)	(3.924)	(3.373)	(1.945)	(0.562)	(2.524)	(2.017)	(1.892)	(3.085)
$log\_filesize$	4.0281***	3.8840***	3.0017***	3.1132***	3.4017***	2.4542***	0.4480	2.8705***	2.5454***	0.8529	1.5331*
	(4.219)	(4.346)	(3.697)	(3.913)	(4.386)	(3.119)	(0.571)	(3.615)	(3.148)	(1.059)	(1.886)
Observations	9910	9905	9905	9904	9903	9901	9898	9895	9896	9897	9895
Δ Observations	8762	8757	8757	8756	8755	8753	8750	8747	8749	8749	8747
Note:											*p<0.1; **p<0.05; ***p<0.01 *(t scores in parentheses)*

Table A2.2: Impact of sentiment and readability on abnormal volatility (2004-2018). Regression is specified as $volatility_{i,t} = \beta\overline{X_{i,t}} + \{year_y\}_{y=2004}^{2018} + \epsilon$, where $X_{i,t}$ takes any of the independent variables. The regressions are performed on each dependent variable individually. Table shows the beta coefficients with its respective t-scores. Each row and column are separate regressions, where the rows represent the independent variable and the columns represent the time horizon of the dependent variable. Δ Observations is number of observations for differenced variables (name starts with "Δ "), while Observations is number of observations for the other regressions.

	Dependent variable: $volume_t$										
	(0 days)	(1 days)	(2 days)	(3 days)	(4 days)	(5 days)	(6 days)	(7 days)	(8 days)	(9 days)	(10 days)
$\Delta naive\_tone$	-0.4797	0.8966	1.0011	-0.5497	-0.0184	0.3958	0.2475	-0.7623	-0.3969	-0.2851	-1.7121*
	(-0.473)	(0.960)	(1.148)	(-0.641)	(-0.022)	(0.451)	(0.290)	(-0.882)	(-0.447)	(-0.312)	(-1.960)
$\Delta tfidf\_tone$	0.1600	0.3360	0.3954	0.0081	0.4331	-0.0774	-0.3880	-0.1565	0.1685	-0.8063	-0.1164
	(0.304)	(0.642)	(0.811)	(0.017)	(0.932)	(-0.163)	(-0.814)	(-0.332)	(0.360)	(-1.612)	(-0.227)
$\Delta vader\_pos$	-0.9736	0.4269	0.5411	0.5838	0.1019	0.2670	-0.6806	-0.3076	-0.0435	-1.2272	-2.3284***
	(-0.920)	(0.443)	(0.641)	(0.703)	(0.119)	(0.302)	(-0.792)	(-0.353)	(-0.050)	(-1.391)	(-2.669)
$\Delta vader\_neg$	-0.5988	-0.5386	-0.8561	-0.9458	-0.0534	-0.3942	-1.1897	-0.7398	-0.8408	-0.2484	0.1334
	(-0.541)	(-0.522)	(-0.948)	(-1.078)	(-0.059)	(-0.427)	(-1.350)	(-0.831)	(-0.923)	(-0.273)	(0.145)
$\Delta vader\_comp$	0.0487	1.0050	0.5482	0.5453	0.0086	-0.0576	0.4783	0.3583	0.7410	-0.3290	-1.3782
	(0.046)	(1.040)	(0.668)	(0.643)	(0.010)	(-0.065)	(0.551)	(0.419)	(0.833)	(-0.376)	(-1.558)
Δnum_sents	0.4756	1.8264*	1.1189	1.7340**	2.0241**	0.8240	0.3085	1.8326**	1.6958*	0.3181	2.0245**
	(0.434)	(1.824)	(1.159)	(2.051)	(2.400)	(1.009)	(0.381)	(2.213)	(1.861)	(0.391)	(2.280)
$gunning\_fog$	1.9023**	2.0818**	1.9499**	3.0471***	2.7823***	1.5570*	0.4431	2.0403**	1.6151**	1.5557*	2.4980***
	(1.972)	(2.370)	(2.473)	(3.924)	(3.373)	(1.945)	(0.562)	(2.524)	(2.017)	(1.892)	(3.085)
$log\_filesize$	4.0281***	3.8840***	3.0017***	3.1132***	3.4017***	2.4542***	0.4480	2.8705***	2.5454***	0.8529	1.5331*
	(4.219)	(4.346)	(3.697)	(3.913)	(4.386)	(3.119)	(0.571)	(3.615)	(3.148)	(1.059)	(1.886)
Observations	9910	9905	9905	9904	9903	9901	9898	9895	9896	9897	9895
Δ Observations	8762	8757	8757	8756	8755	8753	8750	8747	8749	8749	8747

Note: *p<0.1; **p<0.05; ***p<0.01

(t scores in parentheses)

Table A2.3: Impact of sentiment and readability on abnormal trading volume (2004-2018). Regression is specified as $volume_{i,t} = \beta\overline{X_{i,t}} + \{year_y\}_{y=2004}^{2018} + \epsilon$, where $X_{i,t}$ takes any of the independent variables. The regressions are performed on each dependent variable individually. Table shows the beta coefficients with its respective t-scores. Each row and column are separate regressions, where the rows represent the independent variable and the columns represent the time horizon of the dependent variable. Δ Observations is number of observations for differenced variables (name starts with "Δ "), while Observations is number of observations for the other regressions.

A2.2 Results without Year Dummies

Table A2.4 is the same as Table 4.1 in Chapter 4, but without any dummy variables for year.

	Dependent variable: $return_t$								
	(1 days)	(3 days)	(5 days)	(10 days)	(30 days)	(45 days)	(60 days)	(180 days)	(252 days)
$\Delta naive\_tone$	0.0685*	0.0732	0.0554	0.0135	-0.0137	-0.1123	-0.0288	0.3686	0.4219
	(1.925)	(1.583)	(1.059)	(0.206)	(-0.125)	(-0.811)	(-0.173)	(1.197)	(1.154)
$\Delta tfidf\_tone$	0.0210	0.0434	0.0328	-0.0323	-0.0735	-0.2302	-0.1025	0.2949	0.3194
	(0.611)	(0.965)	(0.631)	(-0.488)	(-0.658)	(-1.634)	(-0.607)	(0.943)	(0.837)
$\Delta vader\_pos$	0.0314	0.0504	0.0068	0.0187	0.0289	0.1284	0.2882*	0.4201	0.7384*
	(0.838)	(1.011)	(0.122)	(0.276)	(0.271)	(0.970)	(1.888)	(1.361)	(1.762)
$\Delta vader\_neg$	-0.0360	-0.0461	-0.0104	-0.0233	-0.0496	0.0990	-0.1155	-0.6204*	-0.7312**
	(-0.881)	(-0.856)	(-0.178)	(-0.329)	(-0.474)	(0.719)	(-0.720)	(-1.937)	(-1.990)
$\Delta vader\_comp$	0.0561	0.0571	0.0116	0.0387	0.0872	-0.0267	0.2076	0.6828**	0.8763**
	(1.536)	(1.189)	(0.209)	(0.574)	(0.821)	(-0.192)	(1.300)	(2.210)	(2.407)
Δnum_sents	-0.0321	-0.0712	-0.0884	-0.0792	-0.0415	0.0934	0.1319	-0.3428	-0.2432
	(-0.672)	(-1.159)	(-1.248)	(-0.920)	(-0.303)	(0.577)	(0.747)	(-1.160)	(-0.698)
$gunning\_fog$	0.0019	-0.0061	0.0708	0.0627	-0.0585	0.0022	0.0188	-0.7523***	-0.8408**
	(0.066)	(-0.144)	(1.476)	(1.009)	(-0.603)	(0.019)	(0.137)	(-2.805)	(-2.404)
$log\_filesize$	-0.0560*	-0.0764*	-0.0193	-0.0877	-0.2206**	-0.1612	-0.0489	-0.8647***	-1.2889***
	(-1.699)	(-1.651)	(-0.371)	(-1.321)	(-2.149)	(-1.315)	(-0.345)	(-3.073)	(-3.364)
Observations	9937	9619	9617	9610	9590	9577	9567	9440	9362
Δ Observations	9847	9753	9750	9743	9721	9708	9694	9568	9488

Note: *(t scores in parentheses)* *p<0.1; **p<0.05; ***p<0.01

Table A2.4: Impact of sentiment and readability on abnormal return without year dummies. Regression is specified as $return_{i,t,d} = \beta\overline{X_{i,t}}\epsilon$, where $X_{i,t}$ takes any of the independent variables. The regressions are performed on each dependent variable individually. Table shows the beta coefficients with its respective t-scores. Each row and column are separate regressions, where the rows represent the independent variable and the columns represent the time horizon of the dependent variable. Δ Observations is number of observations for differenced variables (name starts with "Δ "), while Observations is number of observations for the other regressions.

Appendix A3

The Impact of Topics from LDA

A3.1 Topic Names and Keywords

We here present a complete table of topics from our LDA model, the respective topic words, together with the weights on each word.

Topic Name	Keywords
Bank and financial lending	0.087*”bank” + 0.054*”credit” + 0.042*”payment” + 0.036*”account” + 0.030*”fund” + 0.029*”amount” + 0.023*”institution” + 0.019*”deposit” + 0.015*”receivables” + 0.014*”issue” + 0.013*”receivable” + 0.012*”fee” + 0.012*”time” + 0.012*”financial” + 0.011*”transfer
Claims and liabilities	0.073*”claim” + 0.036*”liability” + 0.028*”estimate” + 0.019*”amount” + 0.015*”future” + 0.014*”relate” + 0.014*”settlement” + 0.013*”insurance” + 0.012*”expect” + 0.011*”payment” + 0.011*”asbestos” + 0.011*”material” + 0.010*”loss” + 0.010*”base” + 0.010*”ultimate
Internal control and management	0.077*”control” + 0.070*”registrant” + 0.057*”internal” + 0.051*”report” + 0.045*”financial” + 0.034*”significant” + 0.033*”officer” + 0.031*”material” + 0.026*”design” + 0.020*”affect” + 0.019*”base” + 0.019*”fraud” + 0.018*”disclose” + 0.018*”certify” + 0.017*”recent
Oil and gas	0.038*”oil” + 0.026*”price” + 0.026*”gas” + 0.022*”production” + 0.017*”drilling” + 0.017*”reserve” + 0.015*”natural” + 0.014*”future” + 0.012*”capital” + 0.011*”well” + 0.009*”cost” + 0.009*”development” + 0.009*”operation” + 0.009*”activity” + 0.009*”estimate
Media and entertainment	0.032*”advertising” + 0.025*”content” + 0.022*”programming” + 0.022*”network” + 0.021*”television” + 0.020*”license” + 0.017*”significant” + 0.017*”revenue” + 0.014*”right” + 0.014*”distribution” + 0.013*”medium” + 0.012*”agreement” + 0.012*”business” + 0.011*”video” + 0.011*”segment
Customer services	0.086*”service” + 0.041*”customer” + 0.040*”revenue” + 0.023*”business” + 0.015*”network” + 0.013*”datum” + 0.012*”provide” + 0.012*”include” + 0.011*”change” + 0.011*”solution” + 0.010*”increase” + 0.010*”technology” + 0.009*”user” + 0.009*”software” + 0.009*”use
Corporate structure	0.632*”corporation” + 0.063*”page” + 0.021*”percent” + 0.018*”subsidiary” + 0.013*”trust” + 0.012*”aggregate” + 0.011*”management” + 0.010*”forbearance” + 0.009*”look” + 0.008*”recapture” + 0.008*”state” + 0.008*”reasonably” + 0.007*”dividend” + 0.007*”expect” + 0.006*”probable”

Health care	0.050*”member” + 0.047*”care” + 0.047*”health” + 0.029*”service” + 0.028*”state” + 0.026*”program” + 0.022*”provider” + 0.019*”medical” + 0.016*”payment” + 0.015*”plan” + 0.015*”federal” + 0.015*”include” + 0.013*”government” + 0.010*”provide” + 0.010*”manage
Obligations and agreements	0.030*”obligation” + 0.029*”agreement” + 0.027*”collateral” + 0.022*”party” + 0.021*”guarantor” + 0.018*”guarantee” + 0.015*”right” + 0.013*”secure” + 0.013*”default” + 0.013*”document” + 0.011*”security” + 0.011*”payment” + 0.010*”permit” + 0.008*”respect” + 0.008*”bankruptcy
Capital strucutre	0.073*”stock” + 0.050*”common” + 0.047*”share” + 0.034*”merger” + 0.031*”harm” + 0.028*”price” + 0.027*”transaction” + 0.016*”agreement” + 0.015*”acquisition” + 0.014*”business” + 0.013*”purchase” + 0.013*”stockholder” + 0.011*”right” + 0.011*”cash” + 0.010*”equity
Contracts	0.143*”contract” + 0.031*”government” + 0.031*”customer” + 0.019*”cost” + 0.018*”project” + 0.017*”include” + 0.016*”program” + 0.016*”work” + 0.015*”system” + 0.015*”service” + 0.014*”contractor” + 0.014*”amount” + 0.013*”performance” + 0.013*”award” + 0.012*”estimate
Taxes and financial regulations	0.065*”tax” + 0.028*”law” + 0.016*”income” + 0.016*”business” + 0.015*”subsidiary” + 0.014*”include” + 0.012*”regulation” + 0.012*”subject” + 0.012*”liability” + 0.010*”taxis” + 0.010*”applicable” + 0.010*”result” + 0.010*”share” + 0.010*”asset” + 0.010*”financial
Financial results	0.049*”result” + 0.030*”operation” + 0.026*”financial” + 0.025*”condition” + 0.025*”affect” + 0.022*”business” + 0.021*”change” + 0.019*”impact” + 0.019*”adversely” + 0.018*”future” + 0.015*”increase” + 0.015*”adverse” + 0.015*”cost” + 0.014*”ability” + 0.013*”effect
Dividends	0.088*”stock” + 0.053*”dividend” + 0.044*”share” + 0.038*”prefer” + 0.033*”holder” + 0.021*”director” + 0.017*”preferred” + 0.015*”series” + 0.014*”date” + 0.014*”payment” + 0.013*”class” + 0.012*”redemption” + 0.011*”period” + 0.011*”number” + 0.011*”right
Market growth	0.022*”business” + 0.015*”growth” + 0.012*”year” + 0.012*”financial” + 0.012*”management” + 0.011*”market” + 0.011*”significant” + 0.010*”continue” + 0.009*”global” + 0.008*”system” + 0.008*”information” + 0.008*”also” + 0.008*”impact” + 0.008*”new” + 0.007*”client
Financial statements	0.043*”financial” + 0.034*”statement” + 0.024*”result” + 0.016*”operation” + 0.015*”report” + 0.015*”cash” + 0.012*”forward” + 0.012*”include” + 0.012*”information” + 0.012*”end” + 0.011*”material” + 0.011*”period” + 0.011*”management” + 0.010*”look” + 0.010*”consolidated
Energy and power cost	0.031*”cost” + 0.026*”power” + 0.025*”energy” + 0.018*”rate” + 0.015*”generation” + 0.015*”utility” + 0.014*”customer” + 0.011*”project” + 0.011*”plant” + 0.010*”include” + 0.009*”electric” + 0.009*”southern” + 0.009*”unit” + 0.009*”expect” + 0.009*”construction
Loan and financing	0.024*”loan” + 0.020*”risk” + 0.017*”loss” + 0.016*”interest” + 0.016*”credit” + 0.015*”value” + 0.014*”rate” + 0.013*”market” + 0.013*”security” + 0.011*”significant” + 0.010*”change” + 0.010*”include” + 0.010*”financial” + 0.010*”capital” + 0.010*”asset
Insurance	0.031*”loss” + 0.028*”insurance” + 0.023*”investment” + 0.019*”security” + 0.012*”premium” + 0.012*”policy” + 0.012*”value” + 0.011*”business” + 0.011*”rate” + 0.010*”change” + 0.010*”market” + 0.010*”include” + 0.010*”income” + 0.010*”reserve” + 0.009*”expect
Natural gas	0.071*”gas” + 0.060*”natural” + 0.033*”pipeline” + 0.023*”increase” + 0.021*”commodity” + 0.016*”price” + 0.015*”service” + 0.015*”project” + 0.014*”storage” + 0.012*”result” + 0.012*”contract” + 0.011*”volume” + 0.011*”primarily” + 0.010*”due” + 0.010*”include

Product development	0.049*”product” + 0.015*”result” + 0.015*”business” + 0.015*”customer” + 0.012*”technology” + 0.011*”new” + 0.010*”revenue” + 0.010*”market” + 0.009*”sale” + 0.009*”intellectual” + 0.008*”adversely” + 0.008*”affect” + 0.008*”future” + 0.007*”significant” + 0.007*”property
Environmental cost	0.038*”environmental” + 0.035*”cost” + 0.023*”liability” + 0.021*”site” + 0.017*”material” + 0.017*”remediation” + 0.014*”operation” + 0.011*”estimate” + 0.010*”derivative” + 0.010*”law” + 0.009*”party” + 0.009*”require” + 0.009*”facility” + 0.009*”matter” + 0.009*”regulation
Lawsuits	0.027*”damage” + 0.024*”seek” + 0.021*”action” + 0.021*”file” + 0.018*”plaintiff” + 0.017*”class” + 0.017*”claim” + 0.015*”court” + 0.013*”case” + 0.013*”certain” + 0.012*”lawsuit” + 0.012*”complaint” + 0.012*”include” + 0.012*”allege” + 0.012*”relief
Employee benefits	0.102*”participant” + 0.057*”plan” + 0.020*”employer” + 0.018*”account” + 0.016*”benefit” + 0.015*”section” + 0.015*”make” + 0.014*”year” + 0.013*”payment” + 0.013*”amount” + 0.012*”administrator” + 0.011*”election” + 0.010*”distribution” + 0.010*”date” + 0.010*”eligible
Products	0.092*”product” + 0.053*”customer” + 0.031*”sale” + 0.028*”cost” + 0.025*”price” + 0.023*”material” + 0.022*”supply” + 0.022*”demand” + 0.021*”supplier” + 0.020*”production” + 0.020*”manufacturing” + 0.019*”inventory” + 0.018*”manufacture” + 0.016*”market” + 0.016*”facility
Management incentive programs	0.068*”performance” + 0.054*”award” + 0.031*”share” + 0.031*”unit” + 0.029*”stock” + 0.018*”base” + 0.016*”plan” + 0.016*”grant” + 0.016*”restrict” + 0.014*”vest” + 0.014*”period” + 0.012*”year” + 0.011*”determine” + 0.011*”participant” + 0.010*”term
Regulations	0.020*”plan” + 0.019*”regulatory” + 0.017*”new” + 0.016*”certain” + 0.016*”rule” + 0.016*”issue” + 0.015*”impact” + 0.015*”additional” + 0.013*”order” + 0.012*”require” + 0.012*”propose” + 0.011*”final” + 0.011*”standard” + 0.011*”federal” + 0.010*”program
Borrowing and lending	0.056*”lender” + 0.047*”agent” + 0.036*”borrower” + 0.030*”loan” + 0.024*”administrative” + 0.014*”time” + 0.011*”subsidiary” + 0.011*”amount” + 0.010*”credit” + 0.010*”make” + 0.010*”applicable” + 0.010*”interest” + 0.009*”document” + 0.009*”rate” + 0.009*”case
Management incentive programs 2	0.055*”share” + 0.034*”stock” + 0.030*”option” + 0.024*”plan” + 0.021*”grant” + 0.016*”award” + 0.016*”exercise” + 0.016*”date” + 0.013*”law” + 0.013*”vest” + 0.011*”employment” + 0.011*”transfer” + 0.010*”restrict” + 0.010*”unit” + 0.010*”datum
Mining	0.023*”coal” + 0.021*”mine” + 0.021*”mining” + 0.020*”production” + 0.019*”water” + 0.018*”operation” + 0.017*”price” + 0.016*”cost” + 0.014*”project” + 0.012*”expect” + 0.011*”year” + 0.010*”rail” + 0.010*”sale” + 0.009*”closure” + 0.009*”copper
Pharmaceuticals	0.048*”product” + 0.019*”development” + 0.018*”drug” + 0.018*”patent” + 0.018*”patient” + 0.018*”approval” + 0.018*”regulatory” + 0.017*”clinical” + 0.015*”candidate” + 0.011*”trial” + 0.011*”include” + 0.011*”sale” + 0.010*”program” + 0.009*”healthcare” + 0.009*”party
Transportation	0.052*”vehicle” + 0.044*”fuel” + 0.032*”aircraft” + 0.021*”travel” + 0.020*”price” + 0.019*”american” + 0.017*”ship” + 0.016*”lease” + 0.015*”increase” + 0.014*”program” + 0.014*”purchase” + 0.014*”car” + 0.012*”revenue” + 0.011*”customer” + 0.011*”fleet
Notes and bonds	0.128*”note” + 0.051*”rate” + 0.051*”interest” + 0.044*”date” + 0.034*”amount” + 0.026*”principal” + 0.021*”payment” + 0.019*”day” + 0.018*”period” + 0.015*”applicable” + 0.013*”accrue” + 0.013*”maturity” + 0.013*”price” + 0.012*”time” + 0.012*”senior

Financial plans	0.055*”plan” + 0.043*”benefit” + 0.026*”pension” + 0.024*”expect” + 0.022*”asset” + 0.019*”rate” + 0.018*”return” + 0.016*”cost” + 0.014*”year” + 0.012*”financial” + 0.012*”term” + 0.012*”tax” + 0.011*”obligation” + 0.010*”include” + 0.010*”long
Retail	0.040*”store” + 0.035*”sale” + 0.034*”consumer” + 0.032*”brand” + 0.027*”fiscal” + 0.019*”retail” + 0.012*”customer” + 0.012*”include” + 0.012*”new” + 0.011*”inventory” + 0.010*”lease” + 0.010*”year” + 0.009*”merchandise” + 0.008*”product” + 0.008*”credit
Transactions	0.041*”seller” + 0.037*”agreement” + 0.029*”buyer” + 0.027*”party” + 0.025*”purchaser” + 0.024*”close” + 0.023*”transaction” + 0.016*”closing” + 0.015*”business” + 0.013*”transfer” + 0.011*”asset” + 0.010*”reasonably” + 0.010*”material” + 0.010*”purchase” + 0.010*”contemplate
Executive benefits	0.056*”executive” + 0.032*”employment” + 0.027*”agreement” + 0.020*”termination” + 0.018*”benefit”+ 0.016*”payment” + 0.014*”date” + 0.013*”pay” + 0.011*”provide” + 0.011*”term” + 0.011*”period” + 0.010*”year” + 0.009*”time” + 0.009*”severance” + 0.008*”plan
Legal boilerplate	0.046*”section” + 0.033*”date” + 0.020*”provide” + 0.019*”applicable” + 0.016*”agreement” + 0.016*”time” + 0.016*”respect” + 0.015*”pursuant” + 0.014*”event” + 0.014*”right”+ 0.013*”make” + 0.013*”day” + 0.013*”provision” + 0.012*”otherwise” + 0.012*”payment
Joint ventures	0.045*”investment” + 0.034*”venture” + 0.031*”interest” + 0.029*”joint” + 0.027*”partnership” + 0.026*”partner” + 0.022*”unit” + 0.021*”equity” + 0.018*”distribution” + 0.018*”operate” + 0.015*”income” + 0.015*”market” + 0.015*”fund” + 0.014*”entity” + 0.014*”development
Employee benefits 2	0.383*”employee” + 0.056*”apply” + 0.029*”plan” + 0.019*”employment” + 0.015*”business” + 0.014*”benefit” + 0.014*”time” + 0.014*”law” + 0.013*”director” + 0.012*”fiscal” + 0.010*”information” + 0.010*”consultant” + 0.010*”service” + 0.009*”policy” + 0.009*”claim
Asset value	0.027*”asset” + 0.024*”value” + 0.020*”tax” + 0.019*”estimate” + 0.018*”cash” + 0.015*”future” + 0.014*”significant” + 0.014*”fair” + 0.013*”expect” + 0.013*”flow” + 0.013*”amount” + 0.012*”result” + 0.012*”income” + 0.012*”base” + 0.012*”change
Agreements	0.045*”agreement” + 0.033*”party” + 0.024*”agree” + 0.023*”information” + 0.014*”provide” + 0.013*”include” + 0.011*”affiliate” + 0.011*”use” + 0.011*”request” + 0.011*”datum” + 0.010*”obligation” + 0.009*”notice” + 0.008*”document” + 0.008*”confidential” + 0.008*”third
Financial securities	0.043*”security” + 0.042*”trustee” + 0.030*”holder” + 0.028*”indenture” + 0.021*”agent” + 0.019*”redemption” + 0.015*”issuer” + 0.012*”principal” + 0.012*”interest” + 0.012*”date” + 0.012*”pay” + 0.011*”notice” + 0.010*”dealer” + 0.010*”payment” + 0.010*”transfer
Real estate development	0.039*”development” + 0.027*”hotel” + 0.026*”cost” + 0.026*”property” + 0.025*”construction” + 0.023*”community” + 0.023*”land” + 0.022*”operator” + 0.017*”project” + 0.014*”owner” + 0.012*”include” + 0.012*”rate” + 0.011*”tower” + 0.011*”resident” + 0.011*”revenue
Property lease	0.128*”lease” + 0.112*”property” + 0.050*”real” + 0.044*”estate” + 0.024*”lessee” + 0.021*”rent” + 0.019*”rental” + 0.016*”sale” + 0.016*”cost” + 0.014*”common” + 0.012*”term” + 0.011*”value” + 0.010*”stockholder” + 0.010*”income” + 0.008*”construction
Debt structure	0.030*”facility” + 0.029*”debt” + 0.027*”credit” + 0.023*”cash” + 0.021*”agreement” + 0.021*”term” + 0.014*”amount” + 0.014*”capital” + 0.013*”certain” + 0.012*”interest” + 0.011*”senior” + 0.011*”acquisition” + 0.009*”additional” + 0.009*”obligation” + 0.009*”indebtedness

Tenants	0.130*”tenant” + 0.037*”lease” + 0.037*”landlord” + 0.025*”space” + 0.018*”premise” + 0.017*”building” + 0.015*”rent” + 0.012*”expense” + 0.011*”use” + 0.011*”grace” + 0.010*”cost” + 0.009*”build” + 0.009*”foot” + 0.009*”term” + 0.009*”provide
Financial results 2	0.027*”quarter” + 0.025*”cost” + 0.024*”month” + 0.023*”end” + 0.018*”increase” + 0.018*”result” + 0.018*”expect” + 0.016*”impact” + 0.016*”sale” + 0.016*”tax” + 0.015*”year” + 0.015*”due” + 0.014*”approximately” + 0.014*”primarily” + 0.013*”segment
Board of directors	0.076*”director” + 0.033*”meeting” + 0.030*”board” + 0.024*”stockholder” + 0.023*”time” + 0.020*”person” + 0.020*”officer” + 0.015*”vote” + 0.014*”shareholder” + 0.011*”share” + 0.009*”notice” + 0.009*”annual” + 0.009*”provide” + 0.009*”power” + 0.008*”proxy
Foreign exchange	0.075*”foreign” + 0.071*”currency” + 0.042*”rate” + 0.041*”exchange” + 0.030*”dollar” + 0.028*”contract” + 0.025*”hedge” + 0.020*”business” + 0.016*”transaction” + 0.016*”earning” + 0.016*”canadian” + 0.014*”change” + 0.014*”risk” + 0.014*”impact” + 0.014*”sale

Table A3.1: Topics and keywords weighted by their relative importance

A3.2 Measures of the LDA model

A3.2.1 Perplexity

The formula for perplexity as defined by Blei et al. (2003) is:

$$perplexity(D_{test}) = exp\left\{ - \frac{\sum_{d=1}^{M} logp(W_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{A3.1}$$

where M is the number of documents, $p(w_d)$ is the per-word likelihood for words w_d in document d, and N_d is the number of words in document d. As one can see from the function, perplexity is reduced when the likelihood increases, and thus a lower perplexity score indicates a better performing model. Perplexity is calculated using a hold-out sample of document data. Blei et al. (2003) used 10% of their data in the hold-out sample, and kept 90% of the data in the LDA model training set.

A3.3 Results

A3.3.1 Abonrmal Return

	Dependent variable: $return_t$						
	(1 days)	(5 days)	(10 days)	(30 days)	(60 days)	(180 days)	(252 days)
Bank and financial lending	0.1758	0.2821	0.6476	0.0840	-0.5227	2.308 4	2.3971
	(0.779)	(0.805)	(1.386)	(0.114)	(-0.454)	(1.162)	(1.023)
Claims and liabilities	0.3082	0.1802	0.1369	0.1702	-0.5853	-0.3236	-1.6634
	(1.371)	(0.578)	(0.351)	(0.299)	(-0.704)	(-0.230)	(-1.018)
Internal control and management	0.2900	-0.7840	-1.0296	-2.3073	-0.1358	0.1627	1.6132
	(0.575)	(-1.359)	(-1.252)	(-1.467)	(-0.064)	(0.038)	(0.329)
Oil and gas	-0.0548	-0.0010	-0.2724	-0.7225	-1.1166	-0.4271	-1.363 8
	(-0.293)	(-0.004)	(-0.742)	(-1.193)	(-1.246)	(-0.259)	(-0.620)
Media and entertainment	0.2534	0.2531	-0.0348	-0.5369	-0.8916	0.3389	-0.1898

	(1.053)	(0.763)	(-0.077)	(-0.806)	(-0.936)	(0.144)	(-0.072)
Customer services	-0.0161	-0.0218	0.0836	0.9549	1.1882	2.6249	3.66 20
	(-0.075)	(-0.063)	(0.210)	(1.465)	(1.252)	(1.434)	(1.515)
Corporate structure	-0.0571	-0.2887	-0.4431	-0.7335	-1.4615*	-2 .1005*	-2.2435
	(-0.385)	(-1.232)	(-1.463)	(-1.348)	(-1.785)	(-1.660)	(-1.482)
Health care	0.2844	**0.9981****	**1.3341*****	**1.9051*****	**2.4048****	3.5602*	**4.6153****
	(1.430)	(2.288)	(2.633)	(2.684)	(2.123)	(1.918)	(2.278)
Obligations and agreements	0.8667*	0.5500	0.3805	0.1574	-1.1414	-2.5845	-1.6286
	(1.652)	(0.964)	(0.604)	(0.157)	(-0.852)	(-1.063)	(-0.569)
Capital structure	-0.0938	0.0228	0.2188	0.5938	0.2690	0.2079	1.253 8
	(-0.557)	(0.086)	(0.634)	(0.868)	(0.251)	(0.114)	(0.541)
Contracts	0.1574	0.4002	0.5973	-0.2657	-0.4970	-0.5476	-1.1879
	(0.486)	(0.976)	(1.305)	(-0.352)	(-0.437)	(-0.244)	(-0.472)
Taxes and financial regulations	0.0322	0.0069	-0.1866	-0.0584	0.4861	0.9664	0.6181
	(0.177)	(0.023)	(-0.518)	(-0.087)	(0.565)	(0.638)	(0.334)
Financial results	0.2223	0.3581	0.4301	0.8399	0.9758	0.7654	1.0687
	(1.057)	(1.133)	(1.127)	(1.282)	(0.948)	(0.413)	(0.522)
Dividends	0.1672	0.0196	0.2870	0.7947	2.0139*	3.0334*	1. 9999
	(0.585)	(0.047)	(0.556)	(0.902)	(1.738)	(1.846)	(0.930)
Market growth	0.0382	-0.1906	-0.3009	-0.7805	-0.4498	-2.0580	-2.76 23
	(0.172)	(-0.649)	(-0.681)	(-1.097)	(-0.429)	(-1.158)	(-1.305)
Financial statements	-0.2729*	0.2801	0.4392	-0.0136	-0.7780	0.9 296	0.2818
	(-1.790)	(1.079)	(1.207)	(-0.020)	(-0.810)	(0.504)	(0.157)
Energy and power cost	-0.0639	0.3556	0.4512	0.8400	0.4353	0.0188	0 .3479
	(-0.320)	(1.070)	(1.245)	(1.132)	(0.446)	(0.014)	(0.233)
Loan and financing	-0.3584*	-0.3137	-0.0779	-0.2372	0.0381	-0.7 249	-2.4873
	(-1.652)	(-1.049)	(-0.177)	(-0.407)	(0.044)	(-0.489)	(-1.346)
Insurance	-0.7946	-0.4100	-0.1758	-0.1386	0.0303	1.5631	1.0857
	(-1.402)	(-0.736)	(-0.229)	(-0.166)	(0.032)	(0.852)	(0.450)
Natural gas	-0.1134	0.1561	0.3168	-0.0797	1.4108*	1.2866	0.20 26
	(-0.648)	(0.553)	(0.968)	(-0.144)	(1.647)	(0.907)	(0.125)
Product development	0.3262	0.0875	-0.1340	0.8512	1.6281	2.6386	2.2 009
	(1.256)	(0.223)	(-0.295)	(1.107)	(1.486)	(1.126)	(0.825)
Environmental cost	0.1791	0.4739*	**0.6638****	0.5088	1.7407*	**3.8489*****	2.8729*
	(1.296)	(1.959)	(2.153)	(0.741)	(1.901)	(2.735)	(1.750)
Lawsuits	**-0.4546****	**-0.6056****	-0.4026	-0.3175	0.1704	-0.182 8	-0.1063
	(-2.422)	(-2.106)	(-1.047)	(-0.528)	(0.170)	(-0.105)	(-0.052)
Employee benefits	0.0500	-0.2440	-0.1095	0.6242	0.1318	2.0872	1.7803
	(0.278)	(-0.923)	(-0.311)	(1.203)	(0.160)	(1.506)	(1.065)
Products	**-0.4919****	-0.0472	-0.0339	0.3254	0.1741	1.2438	2.35 27
	(-2.511)	(-0.156)	(-0.087)	(0.483)	(0.175)	(0.710)	(1.186)
Management incentive programs	0.2140	0.0992	0.1158	0.2970	0.5742	1.3 187	3.1372
	(1.069)	(0.279)	(0.261)	(0.452)	(0.602)	(0.559)	(0.993)
Regulations	0.0545	-0.4154	-0.5420	0.0329	-0.3407	-2.8121	-1.4136
	(0.240)	(-1.278)	(-1.492)	(0.058)	(-0.372)	(-1.293)	(-0.569)
Borrowing and lending	-0.2024	0.2408	0.3993	0.4107	0.1903	-0.5298	-0.1397
	(-0.926)	(0.706)	(0.856)	(0.630)	(0.211)	(-0.325)	(-0.062)
Management incentive programs 2	0.1151	0.2453	0.3227	0.9338	1.0075	1 .3275	2.8483
	(0.655)	(0.902)	(0.921)	(1.559)	(1.132)	(0.905)	(1.612)
Mining	0.3744	0.1021	0.5975	1.0209	0.3658	0.7375	3.8568
	(1.165)	(0.245)	(1.118)	(1.235)	(0.310)	(0.318)	(1.279)
Pharmaceuticals	0.0256	0.0661	-0.0314	-0.2628	-0.5078	1.3465	2.0035
	(0.063)	(0.133)	(-0.057)	(-0.342)	(-0.511)	(0.634)	(0.780)
Transportation	0.1073	0.0744	0.3504	0.5584	0.1735	0.1565	1.4134
	(0.530)	(0.257)	(0.876)	(0.829)	(0.145)	(0.082)	(0.648)
Notes and bonds	-0.2572	-0.2189	0.0991	-0.7125	-0.5302	-2.0736	-2.1958
	(-1.609)	(-0.801)	(0.273)	(-0.742)	(-0.426)	(-1.117)	(-0.980)
Financial plans	-0.0195	-0.1159	0.3549	1.1535*	**2.1646****	**4.0484****	**4.8328****
	(-0.098)	(-0.386)	(0.957)	(1.946)	(2.328)	(2.301)	(2.227)
Retail	**0.4756****	0.1704	0.3838	0.9502	0.3202	-0.0228	1.4197
	(2.236)	(0.545)	(1.019)	(1.571)	(0.345)	(-0.015)	(0.766)
Transactions	-0.5190*	-0.6458*	-0.1696	0.0731	0.7819	-0.60 42	0.7015
	(-1.725)	(-1.827)	(-0.354)	(0.109)	(0.698)	(-0.309)	(0.303)
Executive benefits	0.0570	-0.0163	0.3483	-0.0429	-0.0487	-2.1174	- 3.0164
	(0.203)	(-0.041)	(0.782)	(-0.064)	(-0.048)	(-1.146)	(-1.418)
Legal boilerplate	0.1709	-0.1118	-0.7296	-0.0388	1.9824	**5.1394****	4.0139*
	(0.586)	(-0.303)	(-1.584)	(-0.053)	(1.613)	(2.422)	(1.691)
Joint ventures	-0.0514	-0.2381	-0.0174	0.6446	1.1228	0.5442	2.3660

	(-0.315)	(-0.677)	(-0.046)	(1.129)	(1.241)	(0.352)	(1.218)
Employee benefits 2	**0.4914****	0.5426	0.3519	0.2558	0.4627	-0.6965	-0.6597
	(2.035)	(1.528)	(0.695)	(0.356)	(0.409)	(-0.356)	(-0.362)
Asset value	-0.0835	-0.1311	-0.4943	**-1.4277****	-1.1623	-0.9991	-2.3741
	(-0.346)	(-0.350)	(-1.076)	(-1.967)	(-1.180)	(-0.564)	(-1.127)
Agreements	-0.1182	-0.2060	0.1017	0.2926	0.5238	-1.1403	0.8550
	(-0.557)	(-0.597)	(0.232)	(0.406)	(0.497)	(-0.597)	(0.372)
Financial securities	-0.1978	-0.4018	-0.0696	-0.4592	0.2589	-0.9318	-2.0424
	(-0.705)	(-0.983)	(-0.145)	(-0.523)	(0.240)	(-0.512)	(-0.996)
Real estate development	-0.0555	-0.0926	0.6777	0.2034	-1.6672	-0.4929	0.6309
	(-0.238)	(-0.254)	(1.545)	(0.306)	(-1.528)	(-0.262)	(0.300)
Property lease	**-0.6485*****	**-0.7888****	-0.8708*	0.1011	-0.5063	0.4527	0.7673
	(-2.603)	(-2.466)	(-1.957)	(0.146)	(-0.487)	(0.252)	(0.353)
Debt structure	0.2982	0.3580	0.6081	0.2305	1.0426	-1.2425	-3.3873
	(1.116)	(0.758)	(1.019)	(0.264)	(0.844)	(-0.494)	(-1.118)
Tenants	-0.0622	0.0459	-0.7190	-0.9263	0.0753	-0.3465	0.0606
	(-0.158)	(0.095)	(-1.392)	(-1.187)	(0.063)	(-0.146)	(0.022)
Financial results 2	0.1885	0.1296	-0.1956	0.1062	-0.6591	-0.6135	-0.8947
	(0.932)	(0.395)	(-0.430)	(0.157)	(-0.644)	(-0.278)	(-0.362)
Board of directors	0.1944	0.3262	**0.7022****	0.5313	0.0184	**2.7871****	**3.2381****
	(1.201)	(1.327)	(2.108)	(0.983)	(0.024)	(2.018)	(2.005)
Foreign exchange	**-0.3267****	-0.1330	-0.5900*	**-1.5043****	-1.1981	**-3.0527****	-2.3909
	(-2.033)	(-0.521)	(-1.869)	(-2.018)	(-1.248)	(-2.164)	(-1.427)
Observations	9847	9750	9743	9721	9694	9568	9488
Note:				*(t scores in parentheses)* *p<0.1; **p<0.05; ***p<0.01			

Table A3.2: Abnormal return regressed against the change in the portion of the annual report related to each topic. Regression is specified as $return_{i,t,t+d} = \{\beta_n\}_{n=1}^{K}\{topic_{n,t}\}_{n=1}^{K} + \{\beta_y\}_{y=1994}^{2018}\{year_y\}_{y=1994}^{2018} + \epsilon_t$ Variables are standardized, so coefficients can be interpreted as percentage abnormal return per standand deviation of abnormal topic discussion. Coefficients with p-values less than 0.05 are in bold.

A3.3.2 Volume and Volatility

	Dependent variable: $volume_t$						
	(0 days)	(1 days)	(2 days)	(3 days)	(4days)	(5 days)	(10 days)
Bank and financial lending	0.3247	-2.2772	0.7076	3.0154	-1.7611	-2.9267	2 .9617
	(0.104)	(-0.785)	(0.237)	(1.119)	(-0.665)	(-1.109)	(1.025)
Claims and liabilities	-1.1470	3.1952	0.8421	1.3474	0.4922	1.2262	0.0445
	(-0.435)	(1.023)	(0.304)	(0.557)	(0.175)	(0.483)	(0.017)
Internal control and management	3.0571	1.6606	-3.3377	-4.6185	-12.3450	-3.1 055	6.6170
	(0.332)	(0.194)	(-0.450)	(-0.607)	(-1.450)	(-0.448)	(0.900)
Oil and gas	1.8599	6.9510***	4.5611	2.1191	0.4374	-1.9340	-0.3788
	(0.833)	(2.607)	(1.634)	(0.966)	(0.187)	(-0.806)	(-0.171)
Media and entertainment	-0.1436	1.5852	1.3163	-0.6834	-0.1801	1.0545	-2.6 702
	(-0.035)	(0.432)	(0.392)	(-0.215)	(-0.055)	(0.301)	(-0.806)
Customer services	-0.3937	-2.2754	-1.0591	-2.1288	0.2817	-0.3965	-3.0282
	(-0.135)	(-0.787)	(-0.365)	(-0.714)	(0.093)	(-0.142)	(-1.099)
Corporate structure	0.9876	0.3255	-0.3201	1.3929	-0.4306	-2.9326	1.4235
	(0.373)	(0.130)	(-0.123)	(0.539)	(-0.170)	(-1.195)	(0.571)
Health care	4.1496	2.3213	3.2267	2.8284	4.8543*	0.3424	-1.806 4
	(1.434)	(0.886)	(1.098)	(0.995)	(1.704)	(0.129)	(-0.679)
Obligations and agreements	5.8683*	1.8531	3.8185	2.5380	5.6309*	6 .7823**	-4.3821
	(1.749)	(0.572)	(1.220)	(0.894)	(1.848)	(2.271)	(-1.535)
Capital strucutre	-1.0977	-2.3234	0.0642	5.0373	3.7502	4.0722	3.3925
	(-0.397)	(-0.684)	(0.021)	(1.575)	(1.204)	(1.232)	(1.141)
Contracts	2.0769	2.1440	-0.1550	2.3621	3.2486	-1.0988	2.0814
	(0.688)	(0.661)	(-0.051)	(0.800)	(1.080)	(-0.393)	(0.775)
Taxes and financial regulations	-3.5808	-4.4349	-3.2073	-0.6792	-0.5746	-1. 6989	1.8064
	(-1.118)	(-1.327)	(-0.974)	(-0.244)	(-0.200)	(-0.551)	(0.597)
Financial results	2.3189	0.4962	-5.1907*	-3.3738	-3.9252	-3.8286	-0. 9671
	(0.688)	(0.149)	(-1.722)	(-1.153)	(-1.277)	(-1.205)	(-0.303)

Dividends	1.1854	-3.1343	0.6887	-3.2841	-2.9733	-4.4370	0.1328
	(0.358)	(-0.994)	(0.209)	(-1.073)	(-1.039)	(-1.498)	(0.043)
Market growth	2.4831	5.0164	2.2305	1.7436	-0.5656	2.0828	3.7704
	(0.807)	(1.589)	(0.717)	(0.578)	(-0.198)	(0.698)	(1.022)
Financial statements	-1.8864	2.9491	-2.5512	-1.8531	-2.0886	-0.9558	-1.13 64
	(-0.651)	(0.909)	(-0.876)	(-0.658)	(-0.711)	(-0.346)	(-0.329)
Energy and power cost	-2.4480	-1.4030	0.3544	1.0049	0.6854	-3.4644	1.2521
	(-0.891)	(-0.508)	(0.135)	(0.443)	(0.268)	(-1.292)	(0.461)
Loan and financing	-4.0981*	-1.6824	1.4388	-1.7939	-1.0089	-0.0716	- 1.1415
	(-1.780)	(-0.688)	(0.613)	(-0.822)	(-0.499)	(-0.032)	(-0.477)
Insurance	1.7645	-0.4180	2.0889	2.2808	4.7041	3.1030	4.0652
	(0.582)	(-0.133)	(0.687)	(0.693)	(1.463)	(1.034)	(1.214)
Natural gas	0.5178	-1.0581	-2.3761	1.8843	0.5481	0.2858	1.4453
	(0.180)	(-0.388)	(-0.739)	(0.764)	(0.191)	(0.103)	(0.510)
Product development	0.4310	0.7921	-2.8517	2.0295	-1.4796	2.2421	0.1542
	(0.133)	(0.264)	(-0.954)	(0.698)	(-0.452)	(0.739)	(0.049)
Environmental cost	8.0165***	-0.1704	3.8679	4.0332*	2.3119	-2.722 3	1.1208
	(2.961)	(-0.067)	(1.494)	(1.674)	(0.885)	(-1.083)	(0.428)
Lawsuits	0.4574	-1.5651	-3.6273	-5.6718**	-2.2596	-2.5999	-3.5785
	(0.165)	(-0.489)	(-1.268)	(-2.191)	(-0.851)	(-0.907)	(-1.264)
Employee benefits	0.3800	1.4216	3.0292	0.2069	-2.7948	-3.9814*	-5.4633**
	(0.151)	(0.529)	(1.152)	(0.084)	(-1.157)	(-1.660)	(-2.097)
Products	1.8591	4.1867	0.9668	2.3779	-2.4113	1.6862	-1.4358
	(0.701)	(1.491)	(0.367)	(0.873)	(-0.884)	(0.667)	(-0.514)
Management incentive programs	0.4355	-0.0473	1.6949	3.2737	-0.5268	-3.2003	-0.4583
	(0.154)	(-0.017)	(0.631)	(1.289)	(-0.187)	(-1.271)	(-0.181)
Regulations	1.7880	2.1965	-1.5880	0.0341	-1.8296	-2.2635	-2.3223
	(0.652)	(0.685)	(-0.583)	(0.012)	(-0.600)	(-0.669)	(-0.807)
Borrowing and lending	-0.6188	2.2193	1.9964	3.3947	2.6664	3.4020	2.6570
	(-0.235)	(0.825)	(0.726)	(1.243)	(0.974)	(1.291)	(0.979)
Management incentive programs 2	3.2568	0.2715	1.4998	1.0591	1.8503	2.7520	2.7251
	(1.252)	(0.101)	(0.573)	(0.445)	(0.755)	(1.097)	(1.119)
Mining	5.3861**	-3.1594	0.5589	0.2139	-5.1286	-5.5290*	-3.3481
	(1.984)	(-1.102)	(0.200)	(0.079)	(-1.596)	(-1.914)	(-1.153)
Pharmaceuticals	2.2334	2.9203	-1.9127	0.1453	-3.5352	-2.2868	2.3434
	(0.647)	(0.842)	(-0.643)	(0.052)	(-1.337)	(-0.900)	(0.867)
Transportation	1.4837	2.2263	2.7264	0.5142	1.3538	5.4389**	3.8450
	(0.509)	(0.846)	(1.094)	(0.207)	(0.480)	(1.981)	(1.455)
Notes and bonds	-6.3776**	-0.5346	-4.0755	2.6234	-1.6951	-1.0855	0.9212
	(-2.224)	(-0.161)	(-1.278)	(0.814)	(-0.544)	(-0.364)	(0.279)
Financial plans	-7.4246***	-1.7937	2.8212	-1.6454	-1.4739	0.4347	-1.6688
	(-2.733)	(-0.602)	(1.007)	(-0.578)	(-0.591)	(0.161)	(-0.620)
Retail	2.6963	-0.5377	1.2283	2.4116	2.0391	1.9147	0.4615
	(0.981)	(-0.193)	(0.480)	(0.945)	(0.793)	(0.737)	(0.167)
Transactions	-1.5807	-3.4686	-4.0192	-1.1717	-8.0925***	-6.3958**	-1.5436
	(-0.504)	(-1.140)	(-1.228)	(-0.376)	(-2.664)	(-2.069)	(-0.507)
Executive benefits	0.0428	-0.2005	3.4148	3.3609	1.2078	2.7742	-2.2137
	(0.015)	(-0.065)	(1.350)	(1.139)	(0.449)	(1.101)	(-0.685)
Legal boilerplate	0.1464	5.4090	0.4787	0.7385	2.0108	4.5092	5.3370
	(0.043)	(1.468)	(0.139)	(0.233)	(0.591)	(1.495)	(1.569)
Joint ventures	0.1892	0.3901	-1.2666	2.9501	5.0643*	5.8989**	1.5392
	(0.068)	(0.135)	(-0.391)	(1.037)	(1.838)	(2.412)	(0.521)
Employee benefits 2	1.8268	-0.9031	-0.8632	-1.8975	-1.3618	-0.1225	-6.6884**
	(0.622)	(-0.315)	(-0.319)	(-0.688)	(-0.527)	(-0.042)	(-2.213)
Asset value	2.4319	3.2086	-2.5852	3.0306	3.0075	2.3398	5.1042
	(0.760)	(0.921)	(-0.699)	(0.896)	(0.958)	(0.787)	(1.619)
Agreements	2.6204	0.4707	-2.1534	2.9987	-1.5446	3.1721	3.3539
	(0.862)	(0.163)	(-0.752)	(1.061)	(-0.503)	(1.155)	(1.160)
Financial securities	1.7184	4.3696	0.9882	5.3326*	3.2198	4.0102	1.8495
	(0.531)	(1.610)	(0.347)	(1.939)	(1.115)	(1.365)	(0.589)
Real estate development	-3.2423	-4.6843	-0.4548	-3.2000	-0.8118	0.4128	-1.6272
	(-1.237)	(-1.617)	(-0.151)	(-1.130)	(-0.299)	(0.143)	(-0.592)
Property lease	-1.2103	-0.9080	-0.1283	3.2947	0.5837	-0.9424	-3.2206
	(-0.411)	(-0.304)	(-0.045)	(1.204)	(0.206)	(-0.335)	(-1.058)
Debt structure	4.8418	2.6969	2.6950	2.5420	6.4628**	0.7691	0.3926
	(1.483)	(0.839)	(0.789)	(0.792)	(2.082)	(0.265)	(0.122)
Tenants	-6.8481*	-3.6276	-2.3303	0.8804	-4.0165	5.3990	2.4636
	(-1.768)	(-0.944)	(-0.618)	(0.261)	(-1.197)	(1.511)	(0.644)

Financial results 2	0.3173	-3.0119	0.6039	0.0242	-0.9361	-3.8215	0.4823
	(0.105)	(-0.974)	(0.225)	(0.010)	(-0.342)	(-1.414)	(0.174)
Board of directors	0.7494	2.7567	0.0869	-1.6643	-2.7742	-0.5626	0.1300
	(0.305)	(0.991)	(0.033)	(-0.639)	(-1.094)	(-0.222)	(0.050)
Foreign exchange	-0.9796	-3.8429	1.3958	-2.0151	-3.4168	-0.1513	-2.2672
	(-0.368)	(-1.424)	(0.556)	(-0.860)	(-1.318)	(-0.061)	(-0.810)
Observations	9825	9822	9820	9819	9818	9815	9810
Note:					*(t scores in parentheses)*	*p<0.1; **p<0.05; ***p<0.01	

Table A3.3: Abnormal trading volume regressed against the change in the portion of the annual report related to each topic. Regression is specified as $volume_{i,t} = \{\beta_n\}_{n=1}^{K}\{topic_{n,t}\}_{n=1}^{K} + \{\beta_y\}_{y=1994}^{2018}\{year_y\}_{y=1994}^{2018} + \epsilon_t$ Variables are standardized, so coefficients can be interpreted as percentage increase in trading volume per standand deviation of abnormal topic discussion.

	Dependent variable: volatility$_t$						
	(0 days)	(1 days)	(2 days)	(3 days)	(4days)	(5 days)	(10 days)
Bank and financial lending	3.4243	-0.7716	4.0767	6.4656	-2.0038	5.4522	1. 0967
	(0.680)	(-0.142)	(0.802)	(1.238)	(-0.425)	(1.055)	(0.206)
Claims and liabilities	-5.3131	4.7119	2.8116	0.1331	3.9013	-1.2076	-5.471 6
	(-0.968)	(0.800)	(0.554)	(0.029)	(0.814)	(-0.276)	(-1.139)
Internal control and management	4.5405	-3.6397	2.0399	0.2104	-2.3578	-14.55 50	13.3128
	(0.257)	(-0.231)	(0.143)	(0.017)	(-0.197)	(-0.901)	(0.993)
Oil and gas	-0.9800	4.2898	-8.0291*	-3.8864	-3.2944	-6.1068	-7.9005*
	(-0.208)	(0.816)	(-1.790)	(-0.848)	(-0.708)	(-1.414)	(-1.929)
Media and entertainment	-5.2743	-3.8862	-2.4190	-2.0077	-5.6307	-6.0549	- 4.8198
	(-0.669)	(-0.606)	(-0.442)	(-0.344)	(-0.962)	(-1.061)	(-0.835)
Customer services	-1.1305	-5.9773	1.1815	-2.6801	3.8755	-1.3108	-0.1774
	(-0.199)	(-1.103)	(0.227)	(-0.536)	(0.764)	(-0.246)	(-0.037)
Corporate structure	9.5582*	10.4454**	2.2979	3.1084	0.3623	-9.421 5**	-4.4922
	(1.743)	(2.028)	(0.459)	(0.531)	(0.075)	(-2.106)	(-0.868)
Health care	-1.4795	-2.1592	6.3486	-0.7245	2.6358	-9.6622*	-6 .2252
	(-0.254)	(-0.399)	(1.205)	(-0.122)	(0.507)	(-1.890)	(-1.197)
Obligations and agreements	4.0210	-0.5926	5.1417	7.6124	4.3156	10.1939*	-8.1697
	(0.625)	(-0.106)	(0.959)	(1.438)	(0.867)	(1.929)	(-1.518)
Capital structure	-3.2788	-4.4549	3.0295	6.8673	0.6754	2.0814	-3.2020
	(-0.596)	(-0.801)	(0.567)	(1.177)	(0.133)	(0.376)	(-0.558)
Contracts	2.7186	8.6403	-0.8485	-5.1987	-1.4425	-5.8372	-0.6660
	(0.452)	(1.563)	(-0.167)	(-0.963)	(-0.306)	(-1.142)	(-0.139)
Taxes and financial regulations	-11.0702*	-6.5219	-9.3988*	5.9341	- 1.2207	3.2193	2.2797
	(-1.869)	(-1.174)	(-1.855)	(1.171)	(-0.247)	(0.560)	(0.414)
Financial results	6.5916	0.6647	-8.0724	-10.7776**	-6.8093	-4.1345	- 2.6650
	(1.051)	(0.100)	(-1.524)	(-2.102)	(-1.233)	(-0.697)	(-0.474)
Dividends	-0.3579	-0.4272	-0.9391	-5.5252	-4.8039	-13.4297**	-1.4803
	(-0.053)	(-0.062)	(-0.156)	(-0.890)	(-0.749)	(-2.472)	(-0.219)
Market growth	-5.3279	3.8987	8.5706	8.0217	-2.7096	5.5583	3.9886
	(-0.930)	(0.647)	(1.469)	(1.533)	(-0.491)	(0.987)	(0.645)
Financial statements	0.3464	0.8114	-1.3266	4.3660	-3.4936	5.0022	-4.0408
	(0.064)	(0.142)	(-0.234)	(0.860)	(-0.644)	(0.980)	(-0.717)
Energy and power cost	3.1488	-2.0910	-0.3148	-1.4401	1.1457	-5.9630	-1.43 30
	(0.530)	(-0.376)	(-0.065)	(-0.308)	(0.223)	(-1.356)	(-0.312)
Loan and financing	-1.0028	5.0484	1.5639	-2.6987	1.8085	3.5320	1.2676
	(-0.224)	(1.062)	(0.353)	(-0.659)	(0.421)	(0.813)	(0.238)
Insurance	-0.8161	-7.0214	-3.4665	-5.5394	-0.5995	-0.2679	4.0457
	(-0.134)	(-1.127)	(-0.608)	(-0.998)	(-0.127)	(-0.049)	(0.684)
Natural Gas	-0.8961	-0.5600	2.1090	4.3137	-8.0543*	-3.0578	9.3315*
	(-0.171)	(-0.109)	(0.413)	(0.885)	(-1.733)	(-0.643)	(1.890)
Product development	2.3487	5.1909	6.5003	4.6617	-8.9946	8.6065	2.2047
	(0.393)	(0.899)	(1.310)	(1.021)	(-1.581)	(1.567)	(0.418)
Environmental cost	5.6999	2.8217	7.7463*	2.8658	-0.0143	-0.8371	5.65 97
	(1.088)	(0.609)	(1.649)	(0.655)	(-0.003)	(-0.193)	(1.149)
Lawsuits	1.4842	-2.1601	-5.6803	-14.3765***	-7.8608	-5.8008	-9.6829*
	(0.255)	(-0.342)	(-1.152)	(-2.858)	(-1.609)	(-1.132)	(-1.894)
Employee benefits	1.2804	9.1232*	-2.4185	-5.9459	-5.4293	-8.4178*	-5.3654

	(0.253)	(1.786)	(-0.479)	(-1.279)	(-1.261)	(-1.886)	(-1.152)
Products	2.1521	6.9585	-1.3727	3.2985	1.3732	-0.2908	-0.8774
	(0.393)	(1.336)	(-0.289)	(0.712)	(0.271)	(-0.060)	(-0.179)
Management incentive programs	-3.2272	-1.1243	-2.7018	-1.7602	0.6691	-9.462 7*	-5.6002
	(-0.601)	(-0.224)	(-0.531)	(-0.375)	(0.133)	(-1.851)	(-1.009)
Regulations	1.6389	1.7789	-3.6742	6.7069	2.1570	-3.4917	-2.1437
	(0.288)	(0.289)	(-0.715)	(1.309)	(0.414)	(-0.665)	(-0.386)
Borrowing and Lending	8.0713	2.2136	5.7985	3.2213	8.1660*	-0.6370	0. 6033
	(1.635)	(0.416)	(1.247)	(0.704)	(1.740)	(-0.130)	(0.127)
Management incentive programs 2	3.5960	-0.3095	0.5511	2.4465	0.1366	0.8852	2.7387
	(0.676)	(-0.062)	(0.111)	(0.555)	(0.028)	(0.194)	(0.549)
Mining	15.6547**	7.4108	2.4787	7.5590	-0.8125	2.2637	0.2746
	(2.517)	(1.336)	(0.481)	(1.547)	(-0.153)	(0.473)	(0.051)
Pharmaceuticals	-3.6446	7.3079	3.1435	-1.5335	-3.0940	2.1689	-4.9656
	(-0.541)	(1.181)	(0.553)	(-0.320)	(-0.623)	(0.432)	(-0.869)
Transportation	8.8531	-5.7423	2.6723	-2.7223	0.3632	3.6631	0.1907
	(1.583)	(-1.247)	(0.573)	(-0.571)	(0.073)	(0.778)	(0.040)
Notes and bonds	-2.3999	2.3324	-6.2791	1.7843	-4.1429	0.0798	7.0887
	(-0.373)	(0.350)	(-1.166)	(0.309)	(-0.781)	(0.015)	(1.113)
Financial Plans	-5.0478	0.0802	1.3698	0.4032	3.6830	1.3104	-4.9808
	(-0.997)	(0.014)	(0.271)	(0.084)	(0.754)	(0.278)	(-0.999)
Retail	2.5788	0.3608	3.5037	4.8584	2.5256	4.5216	7.2889
	(0.497)	(0.074)	(0.737)	(1.084)	(0.527)	(0.956)	(1.480)
Transactions	4.5791	2.0242	-1.3379	-2.9971	-1.0456	-8.4948	0.0365
	(0.744)	(0.352)	(-0.254)	(-0.551)	(-0.183)	(-1.519)	(0.007)
Executive benefits	-1.5080	-2.7119	1.4943	10.2858**	-3.9638	-2.1888	1.7925
	(-0.254)	(-0.489)	(0.314)	(2.028)	(-0.858)	(-0.470)	(0.361)
Legal boilerplate	3.2697	8.7859	14.1733**	3.9020	7.3920	9.4119	10.5893*
	(0.488)	(1.423)	(2.250)	(0.670)	(1.244)	(1.634)	(1.679)
Joint ventures	8.6764*	-3.0565	2.5164	-0.5182	8.5897*	3.2908	-3.2107
	(1.721)	(-0.581)	(0.492)	(-0.100)	(1.720)	(0.617)	(-0.609)
Employee benefits 2	6.6455	3.7762	3.3786	2.1551	6.8410	7.7522	-1.2263
	(1.156)	(0.713)	(0.699)	(0.498)	(1.434)	(1.635)	(-0.234)
Asset value	9.5276	1.7566	-5.7902	3.5573	3.7783	2.1167	4.0443
	(1.608)	(0.255)	(-0.867)	(0.644)	(0.657)	(0.389)	(0.766)
Agreements	1.7460	4.8538	-1.7791	5.4910	-9.1184*	5.4379	6.5702
	(0.295)	(0.843)	(-0.350)	(1.170)	(-1.675)	(1.111)	(1.317)
Financial securities	0.4922	-5.7037	-8.0023	7.2407	-1.9646	-2.8816	-4.8371
	(0.070)	(-0.989)	(-1.512)	(1.362)	(-0.395)	(-0.574)	(-0.770)
Real estate development	2.5189	3.8709	2.8362	-0.8224	-2.6560	4.1707	-4.3913
	(0.454)	(0.782)	(0.504)	(-0.173)	(-0.524)	(0.895)	(-0.883)
Property lease	-6.3309	-2.9557	-1.9153	4.4179	-1.7222	-0.3280	-3.8329
	(-1.132)	(-0.520)	(-0.392)	(0.943)	(-0.363)	(-0.067)	(-0.682)
Debt structure	5.8231	0.9267	-4.2742	4.4078	15.8249***	10.0850*	0.3555
	(0.975)	(0.168)	(-0.759)	(0.769)	(2.980)	(1.889)	(0.067)
Tenants	8.8769	8.0617	-5.0084	2.7532	-2.4562	12.8903**	5.0050
	(1.269)	(1.220)	(-0.787)	(0.519)	(-0.409)	(2.290)	(0.833)
Financial results 2	-1.7650	-3.6157	6.4750	4.1420	3.9509	-10.7341**	2.4772
	(-0.304)	(-0.617)	(1.174)	(0.823)	(0.767)	(-2.064)	(0.465)
Board of directors	0.2484	5.4786	1.2679	-4.4816	-4.5925	1.2866	1.0085
	(0.050)	(1.101)	(0.279)	(-0.977)	(-1.076)	(0.278)	(0.204)
Foreign exchange	-2.5979	-6.4104	4.7749	-5.5675	-10.6515**	-4.5936	-7.9897
	(-0.515)	(-1.318)	(1.080)	(-1.241)	(-2.304)	(-1.007)	(-1.549)
Observations	9824	9819	9819	9818	9816	9814	9808
Note:				*(t scores in parentheses)*	*p<0.1; **p<0.05; ***p<0.01		

Table A3.4: Abnormal volatility regressed against the change in the portion of the annual report related to each topic. Regression is specified as $volatility_{i,t} = \{\beta_n\}_{n=1}^{K}\{topic_{n,t}\}_{n=1}^{K} + \{\beta_y\}_{y=1994}^{2018}\{year_y\}_{y=1994}^{2018} + \epsilon_t$ Variables are standardized, so coefficients can be interpreted as percentage increase in volatility per standand deviation of abnormal topic discussion.

A3.3.3 Results 2004-2018

	Dependent variable: return$_t$						
	(1 days)	(5 days)	(10 days)	(30 days)	(60 days)	(180 days)	(252 days)
Bank and financial lending	0.4335*	0.4167	0.8745**	1.1760*	0.4273	2.7631	1.7397
	(1.788)	(1.238)	(2.096)	(1.704)	(0.384)	(1.239)	(0.732)
Claims and liabilities	0.3252	0.5492*	0.5257	0.6313	-0.1391	-0. 1884	-0.5639
	(1.363)	(1.839)	(1.432)	(1.238)	(-0.176)	(-0.122)	(-0.330)
Internal control and management	-0.1880	-0.7447	-0.5602	-0.4372	2.2651	2.4605	3.0567
	(-0.268)	(-0.926)	(-0.452)	(-0.227)	(0.817)	(0.460)	(0.521)
Oil and gas	-0.1064	-0.1135	-0.3211	-1.1576*	-1.8383*	-0.3 616	-0.6407
	(-0.534)	(-0.364)	(-0.851)	(-1.852)	(-1.928)	(-0.201)	(-0.265)
Media and entertainment	0.2481	0.3169	0.4044	0.6144	0.4514	2.0750	1.2178
	(1.098)	(1.005)	(1.013)	(0.982)	(0.479)	(0.878)	(0.442)
Customer services	0.0489	0.1170	0.2325	0.9044	1.2261	2.7877	4.0704 *
	(0.222)	(0.354)	(0.612)	(1.416)	(1.337)	(1.612)	(1.730)
Corporate structure	-0.1748	-0.3329	-0.4616	-0.6661	-2.0882***	-1.5837	-2.0402
	(-1.107)	(-1.490)	(-1.547)	(-1.445)	(-2.583)	(-1.170)	(-1.206)
Health care	0.1834	1.0261**	1.4692***	1.9203***	2.3237*	3.8267*	4.5871**
	(0.870)	(2.250)	(2.751)	(2.692)	(1.948)	(1.894)	(2.094)
Obligations and agreements	1.0580**	0.9222	0.6306	-0.4475	-1.2516	-0.9826	0.0502
	(1.971)	(1.643)	(1.023)	(-0.476)	(-0.971)	(-0.399)	(0.017)
Capital strucutre	-0.2764*	-0.2762	-0.0324	0.5971	0.1986	1.2352	2.5037
	(-1.874)	(-1.080)	(-0.096)	(0.961)	(0.187)	(0.675)	(1.040)
Contracts	0.1311	0.2187	0.1050	-0.3063	-0.9803	-2.6539	-2.6370
	(0.359)	(0.487)	(0.228)	(-0.407)	(-0.883)	(-1.155)	(-1.004)
Taxes and financial regulations	0.0071	-0.1448	0.0538	0.6431	0.6812	1.6097	1.7878
	(0.035)	(-0.525)	(0.163)	(1.145)	(0.867)	(1.102)	(0.958)
Financial results	0.1032	0.0769	0.2140	0.7574	1.2127	0.4042	0.3302
	(0.487)	(0.238)	(0.564)	(1.124)	(1.179)	(0.201)	(0.151)
Dividends	0.0002	0.0379	0.4254	1.2380	3.1393**	4.3602***	4.0245*
	(0.001)	(0.085)	(0.731)	(1.348)	(2.472)	(2.722)	(1.953)
Market growth	-0.0306	0.0813	0.0078	-0.5313	-1.1839	-1.8574	-1.914 0
	(-0.176)	(0.275)	(0.022)	(-0.919)	(-1.242)	(-1.049)	(-0.936)
Financial statements	-0.0669	0.1575	0.0606	0.3141	-0.2632	-0.4299	-0.6604
	(-0.466)	(0.699)	(0.180)	(0.493)	(-0.295)	(-0.217)	(-0.391)
Energy and power cost	-0.0237	0.4658	0.5554	1.2710*	0.9085	0.57 72	0.0405
	(-0.108)	(1.340)	(1.546)	(1.736)	(0.937)	(0.439)	(0.026)
Loan and financing	-0.4290*	-0.3031	0.0544	0.0187	0.2248	-0.703 7	-3.1951
	(-1.765)	(-0.902)	(0.111)	(0.029)	(0.232)	(-0.424)	(-1.552)
Insurance	-0.7538	-0.4444	-0.3369	-0.0149	-0.1617	0.4395	0.4312
	(-1.227)	(-0.758)	(-0.421)	(-0.017)	(-0.164)	(0.221)	(0.169)
Natural gas	-0.1980	-0.0837	0.2557	-0.2092	0.9814	-0.0303	-1.7676
	(-1.209)	(-0.311)	(0.849)	(-0.452)	(1.313)	(-0.022)	(-1.071)
Product development	-0.0195	0.0248	-0.2719	0.4785	1.1768	2.4873	-0 .1629
	(-0.094)	(0.085)	(-0.737)	(0.647)	(1.095)	(1.044)	(-0.062)
Environmental cost	0.0836	0.4513*	0.7479**	1.3514**	2.3 549**	5.5116***	5.8876**
	(0.533)	(1.760)	(2.180)	(1.979)	(2.336)	(2.724)	(2.396)
Lawsuits	-0.2407	-0.3002	-0.1965	-0.2476	0.3000	1.3651	1.9886
	(-1.434)	(-1.180)	(-0.612)	(-0.481)	(0.318)	(0.786)	(0.950)
Employee benefits	-0.0391	-0.1212	0.0235	1.0022**	1. 2641	1.6887	1.9365
	(-0.210)	(-0.474)	(0.068)	(1.973)	(1.320)	(0.703)	(0.793)
Products	-0.4052*	-0.0703	0.0613	0.0578	-0.4091	0.2049	-0.03 82
	(-1.906)	(-0.225)	(0.153)	(0.084)	(-0.361)	(0.111)	(-0.018)
Management incentive programs	0.1935	0.2073	0.1093	0.9584	0.9603	1.6 767	3.2256
	(0.972)	(0.604)	(0.253)	(1.524)	(1.050)	(0.703)	(1.005)
Regulations	-0.0340	-0.4892	-0.7601**	0.1876	0.0582	-2.2593	- 1.3924
	(-0.136)	(-1.378)	(-1.984)	(0.335)	(0.061)	(-0.937)	(-0.510)
Borrowing and Lending	-0.3692*	-0.0470	-0.1259	0.4309	0.3802	-1 .0263	-1.2396
	(-1.649)	(-0.140)	(-0.290)	(0.699)	(0.421)	(-0.594)	(-0.500)
Management incentive programs 2	0.0418	0.2703	0.4720	0.6524	0.9682	1 .6023	2.3013
	(0.255)	(1.098)	(1.562)	(1.296)	(1.188)	(1.102)	(1.308)
Mining	0.5545	0.2922	0.9818	1.2153	0.5245	1.8027	3.5637
	(1.429)	(0.594)	(1.607)	(1.363)	(0.411)	(0.732)	(1.071)
Pharmaceuticals	-0.1540	-0.0408	-0.2064	0.0377	0.0321	0.8472	1.545 8
	(-0.793)	(-0.140)	(-0.618)	(0.059)	(0.031)	(0.384)	(0.569)
Transportation	0.1491	0.1556	0.4956	0.8180	0.2358	0.0279	0.3917
	(0.741)	(0.559)	(1.275)	(1.261)	(0.184)	(0.014)	(0.176)

Notes and bonds	-0.0339	-0.1436	-0.0601	-0.0538	1.1053	1.2654	0.54 56
	(-0.195)	(-0.460)	(-0.174)	(-0.054)	(0.893)	(0.745)	(0.274)
Financial plans	-0.0363	-0.1906	-0.1469	0.2172	2.0211**	3.2390*	2.8042
	(-0.182)	(-0.655)	(-0.421)	(0.406)	(2.155)	(1.769)	(1.215)
Retail	0.3780	0.0195	0.0891	0.5705	0.9456	1.0904	3.0660*
	(1.627)	(0.058)	(0.235)	(1.000)	(1.030)	(0.708)	(1.704)
Transactions	-0.4556	-0.3161	-0.0494	-0.0326	1.1617	-0.1315	0.4417
	(-1.376)	(-0.825)	(-0.100)	(-0.049)	(0.980)	(-0.065)	(0.178)
Executive benefits	-0.1080	-0.2710	-0.1865	-1.0459	-0.4321	-1.9522	-2.4704
	(-0.354)	(-0.635)	(-0.391)	(-1.633)	(-0.446)	(-1.117)	(-1.206)
Legal boilerplate	-0.1209	-0.5901	-0.8185*	0.0381	1.1675	3.5872*	2.7588
	(-0.338)	(-1.328)	(-1.731)	(0.053)	(0.903)	(1.676)	(1.160)
Joint ventures	0.0498	-0.2704	-0.0141	0.8175	0.6819	-0.1639	-0.1336
	(0.301)	(-0.854)	(-0.037)	(1.597)	(0.794)	(-0.104)	(-0.064)
Employee benefits 2	0.4914**	0.5426	0.3519	0.2558	0.4627	-0.6965	-0.6597
	(2.035)	(1.528)	(0.695)	(0.356)	(0.409)	(-0.356)	(-0.362)
Asset value	-0.0835	-0.1311	-0.4943	-1.4277**	-1.1623	-0.9991	-2.3741
	(-0.346)	(-0.350)	(-1.076)	(-1.967)	(-1.180)	(-0.564)	(-1.127)
Agreements	-0.1182	-0.2060	0.1017	0.2926	0.5238	-1.1403	0.8550
	(-0.557)	(-0.597)	(0.232)	(0.406)	(0.497)	(-0.597)	(0.372)
Financial securities	-0.1978	-0.4018	-0.0696	-0.4592	0.2589	-0.9318	-2.0424
	(-0.705)	(-0.983)	(-0.145)	(-0.523)	(0.240)	(-0.512)	(-0.996)
Real estate development	-0.0555	-0.0926	0.6777	0.2034	-1.6672	-0.4929	0.6309
	(-0.238)	(-0.254)	(1.545)	(0.306)	(-1.528)	(-0.262)	(0.300)
Property lease	-0.6485***	-0.7888**	-0.8708*	0.1011	-0.5063	0.4527	0.7673
	(-2.603)	(-2.466)	(-1.957)	(0.146)	(-0.487)	(0.252)	(0.353)
Debt structure	0.2982	0.3580	0.6081	0.2305	1.0426	-1.2425	-3.3873
	(1.116)	(0.758)	(1.019)	(0.264)	(0.844)	(-0.494)	(-1.118)
Tenants	-0.0622	0.0459	-0.7190	-0.9263	0.0753	-0.3465	0.0606
	(-0.158)	(0.095)	(-1.392)	(-1.187)	(0.063)	(-0.146)	(0.022)
Financial results 2	0.1885	0.1296	-0.1956	0.1062	-0.6591	-0.6135	-0.8947
	(0.932)	(0.395)	(-0.430)	(0.157)	(-0.644)	(-0.278)	(-0.362)
Board of directors	0.1944	0.3262	0.7022**	0.5313	0.0184	2.7871**	3.2381**
	(1.201)	(1.327)	(2.108)	(0.983)	(0.024)	(2.018)	(2.005)
Foreign exchange	-0.3267**	-0.1330	-0.5900*	-1.5043**	-1.1981	-3.0527**	-2.3909
	(-2.033)	(-0.521)	(-1.869)	(-2.018)	(-1.248)	(-2.164)	(-1.427)
Observations	8784	8695	8688	8668	8646	8530	8455
Note:				*(t scores in parentheses)* *p<0.1; **p<0.05; ***p<0.01			

Table A3.5: Abnormal return regressed against the change in the portion of the annual report related to each topic (2004-2018). Regression is specified as $return_{i,t,t+d} = \{\beta_n\}_{n=1}^{K}\{topic_{n,t}\}_{n=1}^{K} + \{\beta_y\}_{y=1994}^{2018}\{year_y\}_{y=1994}^{2018} + \epsilon_t$ Variables are standardized, so coefficients can be interpreted as percentage abnormal return per standand deviation of abnormal topic discussion.

Appendix A4

Trading

A4.1 The AdaBoost-SAMME Algorithm

Following the terminology of Hastie et al. (2009), we here present the multi-class AdaBoost algorithm (AdaBoost-SAMME):

We are given a set of training data $(\vec{x_1}, c_1), (\vec{x_2}, c_2), ..., (\vec{x_n}, c_n)$ where $\vec{x_i} \in \mathbb{R}^P$, and the output c_i assumes values in a finite set $\{1, 2, ..., K\}$ where K is the number of classes. The goal is to find a classification rule $C(\vec{x})$ from the training data to enable us to use this rule to predict new and unseen instances of $\vec{x}$. AdaBoost works by first building a weak classifiec (e.g., a C4.5 Decision Tree). The weight of the misclassified data points are then increased, before a new tree is buildt, forcing the new classifier to put additional emphasis on the previously misclassified instances. We let $T(\vec{x})$ denote a weak multi-class classifier that maps $\vec{x}$ to classes. The procedure is presented in algorithm 1.

Algorithm 1: AdaBoost-SAMME (multi-class AdaBoost)

```
Initialize the observation weights $w_i = 1/n.$;
for m=1 to M: do
    1. Fit a classifier $T^{(m)}(x)$ to the training data using weights $w_i$;
    2. Compute
           $err^{(m)} = \sum_{i=1}^{n} w_i \mathbb{I}\left(c_i \neq T^{(m)}(\vec{x}_i)\right) / \sum_{i=1}^{n} w_i$
    3. Compute
           $\alpha^{(m)} = log \frac{1 - err^{(m)}}{err^{(m)}} + log(K - 1)$
    4. Set
           $w_i \leftarrow w_i \cdot exp\left(\alpha^{(m)} \cdot \mathbb{I}\left(c_i \neq T^{(m)}(\vec{x}_i)\right)\right),$
       for i= 1, 2,...,n.
    5. Re-normalize $w_i$
end
Result: Output
           $C(\vec{x}) =_k \sum_{m=1}^{M} \alpha^{(m)} \cdot \mathbb{I}(T^{(m)}(\vec{x}) = k)$
```